

# Asymmetric, Closed-Form, Finite-Parameter Models of Multinomial Choice

Timothy Brathwaite<sup>a,\*</sup>, Joan Walker<sup>b</sup>

<sup>a</sup>*Department of Civil and Environmental Engineering, University of California at Berkeley  
116 McLaughlin Hall, University of California, Berkeley, CA, 94720-1720*

<sup>b</sup>*Department of Civil and Environmental Engineering, University of California at Berkeley  
111 McLaughlin Hall, University of California, Berkeley, CA, 94720-1720*

---

## Abstract

In transportation, the number of observations associated with one discrete outcome is often greatly different from the number of observations associated with another discrete outcome. This situation is known as class-imbalance. Consider the choice of commute travel mode. In the United States, there are almost always many more automobile commuters than bicycle commuters. Using typical discrete choice models, where a linear-in-parameters index  $V$  is used to represent the “systematic utility” of an alternative, such class-imbalance would usually be explained by the magnitudes of the indices. The index of the under-represented alternative would generally be much lower than the index of the over-represented alternative. However, one might alternatively hypothesize that class imbalance comes not from much lower index values of the under-represented alternative relative to the over-represented alternative but instead from unequal (i.e. asymmetric) sensitivities to low and high values of the under-represented alternative’s index. That is, one may hypothesize that class imbalance is a product of an asymmetric probability function. This hypothesis implies that the probability of choosing the under-represented alternative decreases more rapidly from 50% than it increases, even for equal-magnitude decreases and increases in the under-represented alternative’s index. Despite being a valid hypothesis for class-imbalanced choice situations, few relatively simple models exist for testing this hypothesis in a multinomial setting. Our paper fills this gap.

In particular, we addressed the following questions: “how can one construct asymmetric, closed-form, finite-parameter models of multinomial choice” and “how do such models compare against commonly used symmetric models?” In answering these questions, we (1) introduced a new class of closed-form, finite-parameter, multinomial choice models that we call “logit-type models,” (2) introduced a procedure for using our logit-type models to extend existing binary choice models to the multinomial setting, and (3) introduced a procedure for creating new binary choice models (both symmetric and asymmetric). Together, our contributions allow us to create new asymmetric, closed-form, finite-parameter multinomial choice models either by creating multinomial extensions of existing, asymmetric, binary choice models or by creating new asymmetric, binary choice models and extending those to the multinomial setting.

As proofs of concept, we developed four new asymmetric, multinomial choice models. Two of these models (the uneven logit model and asymmetric logit model) are completely new introductions to the literature that we created. The other two models (the complementary log-log model and the scobit model) are asymmetric choice models that we generalized from the binary to the multinomial setting for the first time. To assess the statistical and practical impacts of our models, we conducted two applications using mode choice models estimated on a dataset of commute trips made by residents of the San Francisco Bay Area in 2012. In each application, we used the same model specifications. For the first application, we judged the impacts on automobile commute mode shares of a congestion charge in Downtown San Francisco using each of our asymmetric models and the multinomial logit (MNL) model as an example of a symmetric choice model. In the second application, we judged the effect of using our asymmetric models versus the MNL model for target selection in an individualized, travel demand management (TDM), marketing campaign.

Overall, we found evidence that the probability functions for our mode choice model should be asymmetric. Most of our asymmetric models statistically outperformed the MNL model. Three out of four of our asymmetric models had much higher in-sample log-likelihoods and out-of-sample log-likelihoods (as

---

\*Corresponding Author

Email addresses: [timothyb0912@berkeley.edu](mailto:timothyb0912@berkeley.edu) (Timothy Brathwaite), [joanwalker@berkeley.edu](mailto:joanwalker@berkeley.edu) (Joan Walker)

judged by 10-fold, stratified cross-validation) than the MNL model. These dominant performances were corroborated and shown to be statistically significant by likelihood-ratio tests that accounted for the additional parameters being estimated. Beyond statistical fit, we also found practical differences between our asymmetric models and the MNL model. In our congestion pricing application, we found that (relative to the asymmetric choice models) the MNL model provided counter-intuitive inferences by over-predicting the amount of drive-transit-walk trips that would take place in San Francisco as opposed to the East Bay. Moreover, in our TDM application, we found that the MNL is fiscally inefficient when compared to the best performing asymmetric choice models, requiring \$40 to \$50 more dollars to attract each new transit rider when the budget for the marketing campaign is low. In summary, our results suggest that while asymmetric models may not always outperform symmetric ones, asymmetric choice models are worth testing in one’s analysis because they might have better statistical performance and entail substantively different policy and financial implications when compared with traditional symmetric models, such as the MNL.

*Keywords:* Asymmetric Probability Function, Parametric Link Function, Discrete Choice Model, Multinomial Logit Model, Closed-form, Class Imbalance

---

## 1. Introduction

Discrete choice modeling is widely used in transportation. It is used in every area of travel demand analysis, such as residential choice, work location choice, destination choice, time-of-travel choice, mode choice, and route choice. Moreover, discrete choice modeling is also used outside of transportation in fields such as marketing, economics, finance, operations research, statistics, and medicine. Across these many disciplines, the most commonly used models have fairly simple functional forms, such as the multinomial logit (MNL) and binary logit models. The use of simple models is, in part, due to the greater computational burdens required to estimate and forecast with very general discrete choice models. Clearly then, it is important to create simple models that are nonetheless able to avoid unwanted properties of classic models such as the MNL model. In this paper, we introduce models that have the same basic form as the MNL model but, for the price of a finite number of new parameters that are to be estimated from the data, provide potentially much better fits to one’s data and avoid a “symmetry property” that we argue is often undesirable. The next paragraph will review the MNL model because it is the starting point for the class of models that we introduce. Then, we will describe the symmetry property, show it is present in common discrete choice models, and make the case that such a property is not always desirable.

While the MNL and binary logit models are often used because of their ease of estimation, their closed-form probability equations (shown in Equation 1)<sup>1</sup>, and their ease of interpretation, their use requires analysts to accept a set of properties that may be overly restrictive or inaccurate in the specific contexts being modeled. Specifically, one well known property is known as Independence from Irrelevant Alternatives (I.I.A.). The I.I.A. property is seen as problematic when one considers substitution patterns between alternatives that are closely related, and there have been numerous models that aim to avoid I.I.A. (e.g. nested logit, cross-nested

---

<sup>1</sup>Note that the variables are fully listed in order to make clear the notation used in the paper.

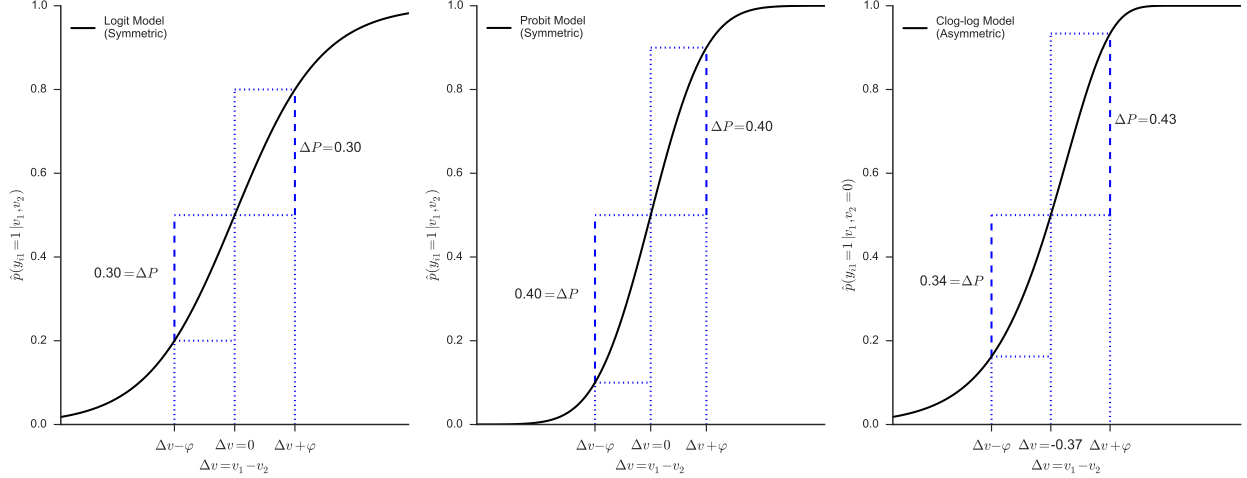


Figure 1: Symmetric and Asymmetric Binary Probability Functions

logit, etc.).

$$P(y_{ij} = 1 | V_{i1}, V_{i2}, \dots, V_{ik} \ \forall \{j, k\} \in C_i) = \frac{\exp(V_{ij})}{\sum_{\ell \in C_i} \exp(V_{i\ell})}$$

where  $y_{ij}$  = a binary (0 or 1) indicator of whether individual  $i$  is associated with outcome  $j$ .

$C_i$  = the choice set for individual  $i$

$V_{ij} = x_{ij}\beta$  = the index for alternative  $j$  for individual  $i$  (1)

$\beta$  = a column vector of unknown population parameters

$x_{ij} = h(z_j, \zeta_i)$ , a row vector.

$h()$  = a function that returns a row vector

$z_j$  = attributes of alternative  $j$  for individual  $i$

$\zeta_i$  = characteristics of individual  $i$

In addition to the I.I.A. property, the MNL model's probability function also implies a "symmetry property." Specifically, from a point where an individual has a 50% probability of choosing an alternative  $j$ , this probability will increase and decrease at equal rates with respect to equal-magnitude increases and decreases in alternative  $j$ 's index,  $V_{ij}$ . Probability functions with this quality are henceforth referred to as symmetric, and probability functions without this property are henceforth referred to as asymmetric. See Figure 1 for a visual depiction of symmetric and asymmetric probability functions. The binary, complementary log-log model (henceforth clog-log model) is described in Section 3.3.1 and used in Figure 1 as an example of an asymmetric probability function. In contrast, the binary logit and binary probit models are used as examples of symmetric probability functions. Note that the logit model is not the only model with the symmetry property. The other commonly used discrete choice model, the simple probit model<sup>2</sup>, is also symmetric. The point being made here is that while it is seldom spoken of, a basic property of standard discrete choice models is that one's probability of choosing a given alternative is symmetric about 50%, with respect to the index,  $V_{ij}$ , of that alternative.

Although models exhibiting the symmetry property are pervasive in discrete choice modeling, there are situations where such a property may seem overly restrictive. Class-imbalanced choice contexts, where the numbers of observations choosing each alternative are unequal, are one such set of situations. Note that in

<sup>2</sup>The 'simple' probit model assumes that the error terms of the utility of each alternative are independent and identically distributed.

transportation, class-imbalanced choice contexts are ubiquitous. For example, in the United States (US), there are almost always many more automobile drivers than bicyclists when modeling commute mode choices. In class-imbalanced situations, it might be natural to hypothesize that the probability of choosing the under-represented alternative decreases more rapidly from 50% than it increases, even for equal-magnitude decreases and increases in the alternative’s index,  $V_{ij}$ . Of course, this hypothesis is not the only plausible explanation for the observed class imbalance. The point, however, is that symmetric probability models prohibit one from investigating hypotheses about the magnitude of changes in the probability of choosing an alternative, from a probability of 50%, with respect to equal-magnitude increases and decreases in that alternative’s index. This is because symmetric probability models assume *a-priori* that the changes in probability are equal.

In light of this undesired symmetry property of common discrete choice models such as the standard MNL and simple probit model, this paper’s contributions to the transportation and discrete choice literature are that it:

1. introduces a general class of closed-form, finite-parameter models for multinomial choice situations that do not necessarily imply symmetric probability functions (as well as four new models within that class),
2. introduces and demonstrates a methodology for
  - (a) extending existing, binary choice models to the multinomial setting and
  - (b) creating new binary choice models (both asymmetric and symmetric),
3. demonstrates that asymmetric probability models can substantially improve upon the fit of standard discrete choice models such as the MNL model, and
4. shows that, compared to symmetric models such as the MNL model, asymmetric probability functions can lead to substantive differences (both quantitatively and qualitatively) in one’s resulting statistical inference and policy-analyses.

The rest of the paper is organized as follows. Section 2 will review related literature and current approaches to producing discrete choice models that are not necessarily symmetric. Section 3 will detail our proposed class of choice models, relate it to the existing literature, and show how one might create such models. Section 4 will describe the estimation procedures for our proposed models, and Section 5 will detail our empirical examples and case studies, comparing our proposed models to existing ones such as the MNL model. Section 6 will discuss extensions of our work and Section 7 will conclude.

## 2. Literature Review

One can partition the asymmetric discrete choice models that have been proposed in the literature based on whether they:

- are binary or multinomial choice models,
- are closed- or open-form<sup>3</sup> models,
- have a null, finite, or infinite<sup>4</sup> set of shape parameters—i.e. parameters that control the shape of the resulting probability function.

To review the literature that this paper builds upon, we will iterate through each of these descriptors in the coming paragraphs—describing the work that has been done so far, how that work relates to or has been used in transportation, and issues with the existing literature that our paper addresses.

First, virtually all research that explicitly focuses on the development of asymmetric choice models has been carried out in the binary setting. Since at least 1976, statisticians and computer scientists have been introducing closed-form, asymmetric generalizations of the standard binary logit model through the use of one or two shape parameters (Prentice, 1976; Pregibon, 1980; Aranda-Ordaz, 1981; Guerrero and Johnson,

---

<sup>3</sup>The probability equation of open-form models contain analytically intractable integrals or infinite sums.

<sup>4</sup>Models with an infinite number of parameters are known as non-parametric or semi-parametric models.

1982; Stukel, 1988; Czado, 1992, 1994; Nagler, 1994; Chen et al., 1999; Vijverberg, 2000; Masnadi-shirazi and Vasconcelos, 2010; Vijverberg and Vijverberg, 2012; Nakayama and Chikaraishi, 2015; Komori et al., 2015). These shape parameters allow one to adapt the shape of the resulting probability function to fit the data at hand. Beyond generalizations of the logit model, a number of binary, asymmetric models that do not nest the logit model have also been proposed in the statistics literature. For example, the clog-log model has been around since at least the 1920s (Fisher, 1922; Yates, 1955; McCullagh and Nelder, 1989), and the GEV regression model (not to be confused with McFadden’s GEV distribution) is a generalization of the clog-log model with one shape-parameter (Wang and Dey, 2010; Calabrese and Osmetti, 2013). Still other asymmetric models of binary choice have been introduced based on skewed normal distributions (Bazán et al., 2010), skewed student’s  $t$ -distributions (Kim, 2002; Kim et al., 2008), and symmetric power distributions (Jiang et al., 2013). Broadly, the binary choice models with one or more shape parameters have been referred to as “parametric link functions” in the statistics literature (McCullagh and Nelder, 1989), and many examples of asymmetric choice models can be found by searching for scholarly articles that use such phrases.

Two problems exist with the asymmetric choice models just discussed. First, while the litany of binary, asymmetric choice models that has been developed may be quite useful, they must be extended to the multinomial setting for use in transportation contexts—contexts where the choice situations are often inherently multinomial. Secondly, the proliferation of binary, asymmetric choice models suggests that no single asymmetric model fits the needs of all researchers. However, no guidance on how to create such asymmetric models has been offered in the literature. The various models cited in the last paragraph were almost all introduced without any explanation of where the functional form for the model came from. Section 3 resolves these issues by detailing a method for extending binary, asymmetric models to the multinomial setting and by introducing a methodology for creating binary, asymmetric choice models.

In addition to all of the binary, closed-form, asymmetric choice models described above, many binary, asymmetric choice models with open-form probability functions have also been proposed. These open-form models are typically one of two major varieties. One type of binary, open-form, asymmetric model uses an asymmetric probability density function for the difference in the error terms of the utilities of the two alternatives. Examples of this type include the aforementioned models that were based on the skewed normal distributions (Bazán et al., 2010) and skewed student’s  $t$ -distributions (Kim, 2002; Kim et al., 2008). The second type of binary, open-form, asymmetric model is based on a mixed logit or mixed probit approach, whereby a random variable with an asymmetric probability density function is added to the index,  $V_{ij}$ , of the alternative of interest. In this second type of model, the random variable with an asymmetric probability density function is multiplied by an unknown coefficient whose value is to be estimated. If the estimated coefficient’s value is zero, then the model reduces to the symmetric probability model (e.g. logit or probit) being used as the kernel of the asymmetric model. Examples of this type of model include the bayesian asymmetric logit and bayesian asymmetric probit models (Chen et al., 1999).

The first type of binary, open-form, asymmetric model described above might be easily extended to the multinomial setting, provided that there exist multivariate versions of the asymmetric probability density functions that are used in the binary case, or provided that such multivariate distributions can be created. This remains an open question. On the other hand, the second type of binary, open-form, asymmetric model can be easily extended to the multinomial setting by simply adding random variables with asymmetric probability density functions to each of the utility functions for the alternatives in one’s model. However, regardless of whether such models can be extended to handle multinomial choice situations, such open-form models will still entail computational burdens in estimation, storage, and forecasting, relative to their closed-form counterparts. In this paper, we focus on developing closed-form, asymmetric choice models because they are less computationally burdensome and more closely parallel the discrete choice models that have been used in all industries (namely closed-form models such as the MNL model).

Regarding the number of shape parameters in one’s model, all of the asymmetric choice models that have been mentioned so far have had one or two shape parameters, with the exception of the binary clog-log model. In contrast to this, many multinomial, asymmetric choice models have been inadvertently<sup>5</sup> created by transportation researchers, and most of them have no shape parameters. Unlike the binary, closed-form,

---

<sup>5</sup>We say inadvertently because in none of the cases cited was the purpose of creating the model to avoid the symmetry property discussed in Section 1.

asymmetric models discussed above—where the functional form of  $P(y_{i1} = 1 \mid V_{i1}, V_{i2})$  is assumed outright—the multinomial, asymmetric models created by transportation researchers come from assuming various distributions for the error terms in the utility equations for each alternative. In particular, multinomial, asymmetric choice models have been derived by transportation researchers by assuming Weibull (Castillo et al., 2008; Fosgerau and Bierlaire, 2009), Rayleigh (Li, 2011), Type II Generalized Logistic (Li, 2011), Pareto (Li, 2011; Mattsson et al., 2014), Exponential (Li, 2011), and Fréchet (Mattsson et al., 2014) distributions for the utilities of one’s alternatives. In each of these cases, the resulting multinomial choice model is asymmetric. A more recent paper (Nakayama and Chikaraishi, 2015) uses a “ $q$ -GEV” distribution for the utility of each alternative, and derives a multinomial, asymmetric choice model with one shape parameter,  $q$ . As a brief aside, Li (2011) and Nakayama and Chikaraishi (2015) note that all of the models described in this paragraph have probability equations with the same functional form as the MNL model, except that  $V_{ij}$  is replaced with  $S_{ij}$ . Here,  $S_{ij} = S(V_{ij}, \gamma_j)$  or  $S_{ij} = S(V_{ij})$  depending on whether the model has shape parameters ( $\gamma_j$ ), and  $S(\cdot)$  is a monotonically increasing function of  $V_{ij}$ . Note also, that the one multinomial, closed-form, asymmetric model that has been introduced in the statistics literature (Das and Mukhopadhyay, 2014) also has this form, except that  $\gamma_j = [\gamma_{j1}, \gamma_{j2}]^T$ , i.e. there are two shape parameters per alternative. This functional form will be mentioned again in Section 3 as it is very similar to the one that we propose in this paper.

While all of the asymmetric models introduced by transportation researchers share the virtue of being able to handle multinomial choice situations, they all share a key drawback: they are only valid for certain values of the index,  $V_{ij}$ . To be concrete, the weibit model of Castillo et al. (2008) and Fosgerau and Bierlaire (2009) is only defined for values of  $V_{ij}$  that are negative. The same is true for utility maximizing models based on the Rayleigh, Type II Generalized Logistic, or Exponential distributions (Li, 2011). If the model is based on the Pareto distribution, then  $V_{ij}$  must be less than negative one (Li, 2011; Mattsson et al., 2014), and if the model is based on the  $q$ -GEV distribution, then  $V_{ij}$  must be greater than or equal to  $\frac{-1}{q-1}$  for whatever value of  $q$  is specified or estimated from one’s data (Nakayama and Chikaraishi, 2015). In choice situations where the index,  $V_{ij}$ , should be comprised of both variables that increase an alternative’s probability of being chosen and variables that decrease an alternative’s probability of being chosen, it can be hard or impossible to meet such constraints on the index’s value or sign. Because of this, the models mentioned in the last paragraph are only applicable in a restrictive set of circumstances. In Section 3, we will introduce a class of multinomial, closed-form, asymmetric choice models that is (in general) free from the sign and magnitude restrictions on  $V_{ij}$  that have limited the usefulness of asymmetric choice models in transportation so far. Our class of models will be shown to include the previously derived models as special cases.

Lastly, transportation researchers in the multinomial setting (Li, 2011), and econometricians in the binary setting (Horowitz, 1993), have specified closed-form, asymmetric choice models that have an infinite number of shape parameters. That is to say, closed-form, asymmetric models have been specified where the function  $S(V_{ij})$ , as defined above, has been estimated non-parametrically. These models are known in the econometrics literature as single-index models (Härdle et al., 1997; Horowitz, 2010). As shown by Li (2011), single-index models can take on symmetric or asymmetric forms. While these models are quite general, and they avoid the problems that come from mis-specifying one’s probability function (Czado and Santner, 1992a; Koenker and Yoon, 2009), they can be difficult to estimate and require rather large sample sizes to estimate with decent precision. For these reasons, we develop a class of models in Section 3 that depends on a finite number of shape parameters, making the class more flexible than the fixed shape models that are classically used in transportation such as MNL models but less computationally burdensome than the single-index models described above.

### 2.1. Summary

Overall, across a variety of academic disciplines, many asymmetric choice models have been created thus far. However, this development has been fragmented and leaves much room for improvement. In particular, most of the existing asymmetric models are binary models, but to be most useful in transportation, these binary models need to be extended to the multinomial setting. Moreover, we need systematic methods for creating new asymmetric models when the existing one’s do not meet our research needs. In the previous literature, there has been much work on creating asymmetric, open-form, binary choice models. In this paper, we do not pursue the development of such models because of their greater computational complexity in

estimation, storage, and forecasting in comparison to their closed-form counterparts. For the same reason, we do not consider closed-form, multinomial, asymmetric models with an infinite number of shape parameters. Instead, we build on the work of transportation researchers since they have created numerous multinomial, asymmetric models that have zero or a finite number of shape parameters. A major limitation of the asymmetric models in transportation is that they all restrict the values that the index,  $V_{ij}$ , can take. In the next section, we address each of these issues by proposing a class of multinomial, closed-form models with zero or a finite number of shape parameters. The proposed class will be able to avoid the symmetry property without restrictions on the index, and it will include many of the existing models as special cases. We will also provide guidance on extending existing binary models to the multinomial setting and on creating new asymmetric choice models.

### 3. A General Class of Asymmetric Models

In this section, we present a class of discrete choice models that can avoid the symmetry property described in Section 1 without imposing restrictions on the sign or magnitude of the index,  $V_{ij}$ , for any given alternative  $j$ . We will proceed as follows. Section 3.1 will give the general formulation of our proposed class of models and show how this formulation can avoid the symmetry property described above. Section 3.2 will then relate our models to existing literature. Next, in Section 3.3 we will demonstrate how our proposed class of models can be used to extend existing, asymmetric, closed-form models of binary choice to the multinomial setting. To do so, we will extend the clog-log model and the scobit model from the binary to the multinomial setting for the first time. Finally, in Section 3.4 we will propose and demonstrate one possible approach to deriving new asymmetric choice models when existing models are not adequate for one's needs. In doing so, we will derive two new asymmetric choice models, the "uneven logit model" and the "asymmetric logit model."

#### 3.1. General Formulation

Our proposed class of models, described below, is appropriate for multinomial choice situations, has a closed-form probability equation, and only depends on a finite number of parameters. Moreover, we refer to our proposed model class as "Logit-Type" models because their probability functions share the same functional form as the MNL model: an exponential term divided by a sum of exponential terms. The probability function for our proposed "logit-type" models is:

$$P(y_{ij} = 1 | \tau, \gamma, V_{i1}, V_{i2}, \dots, V_{ik} \ \forall \{j, k\} \in C_i) = \frac{\exp[\tau_j + S(V_{ij}, \gamma_j)]}{\sum_{\ell \in C_i} \exp[\tau_\ell + S(V_{i\ell}, \gamma_\ell)]}$$

$$= \frac{\exp(S_{ij})}{\sum_{\ell \in C_i} \exp(S_{i\ell})}$$

where  $\tau$  = a 1-dimensional vector of constants, with one value for each alternative in the dataset.

$\gamma$  = a 2-dimensional matrix of shape parameters, with one column for each alternative in the dataset. (2)

$\tau_j$  = a constant associated with alternative  $j$ .

$\gamma_j$  = a column vector of shape parameters associated with alternative  $j$ .

$S(\cdot)$  = a closed-form, model-specific function of  $V_{ij}$  and  $\gamma_j$ . It is monotonically increasing in  $V_{ij}$ .

As before, if a model has no shape parameters, then we replace  $S(V_{ij}, \gamma_j)$  with  $S(V_{ij})$ .

Note that unlike standard logit models, our class of logit-type models makes no assumptions regarding additive random utility functions. As shown by Mattsson et al. (2014), models of the form given in Equation 2 are obtainable under an infinite number of random utility specifications, not all of which are additive. One example we have already mentioned is the case of multiplicative utilities (Castillo et al., 2008; Fosgerau and

Bierlaire, 2009) that are Weibull distributed. As shown by this example, it may be incorrect to interpret  $S_{ij} = \tau_j + S(V_{ij}, \gamma_j)$  as a redefinition of the systematic portion of one’s utility. Expressions such as  $S_{ij}$  may arise solely due to the derivation of the choice probabilities, and they may not actually be present in the simplest expression of the random utility.

To see how our proposed model class can avoid the symmetry property described in the introduction, one can make the analogy between the logit-type models given in Equation 2 and the MNL model given in Equation 1. Since the two models share the exact same functional form except for the replacement of  $V_{ij}$  with  $S_{ij}$  and since variable names do not influence mathematical properties, it follows that logit-type models are symmetric with respect to  $S_{ij}$ . This means that for logit-type models, equal-magnitude increases and decreases in the probability of choosing alternative  $j$  (from an initial probability of 50%) will result if and only if equal-magnitude increases and decreases in  $S_{ij}$  are experienced. As a consequence, one will avoid the aforementioned symmetry property if and only if equal-magnitude increases and decreases in  $V_{ij}$  (respectively) lead to unequal increases and decreases in  $S_{ij}$ . Formally, if

$$S(V_{ij} + \varphi, \gamma_j) - S(V_{ij}, \gamma_j) \neq S(V_{ij}, \gamma_j) - S(V_{ij} - \varphi, \gamma_j), \forall \varphi \geq 0 \quad (3)$$

then the logit-type model given by Equation 2 will be asymmetric with respect to  $V_{ij}$ .

### 3.2. Relation to existing literature

The logit-type models described above encapsulate and generalize many closed-form, finite-parameter, discrete choice models that exist in the literature. For example, one can use Equation 2 to denote the models described by Li (2011) that were based on assuming Weibull, Rayleigh, Type II Generalized Logistic, Pareto, or Exponential distributions for one’s utilities. Except for the model based on the weibull distribution,  $\tau_j = 0 \forall j$  and there are no shape parameters. When basing one’s choice model on Weibull distributed utilities, we have  $\gamma_j = \gamma^* \forall j$  where  $\gamma^*$  is the scale parameter of the distributions. The precise transformations,  $S(\cdot)$ , are provided in Table 2 of Li (2011) for each distribution mentioned above. Likewise the asymmetric, closed-form, multinomial choice model of Das and Mukhopadhyay (2014) is a special case of the models given in Equation 2. Here, again,  $\tau_j = 0 \forall j$ . However, their model has two shape parameters for each alternative, so  $\gamma$  has two rows, and  $S(\cdot)$  is now given by their function  $G(\cdot)$  (Das and Mukhopadhyay, 2014). Examples of binary models that are special cases of logit-type models will be given in Section 3.3 when we demonstrate how one can use Equation 2 to extend existing binary models to the multinomial setting.

To be clear, not all closed-form, finite-parameter discrete choice models are special cases of logit-type models. An example of a closed-form, finite-parameter, multinomial model that is not a special case of a logit-type model is the “exponential choice model” (also known as the “negative exponential distribution” model) (Daganzo, 1979; Alptekinoglu and Semple, 2016). This can be most easily seen by considering the fact that logit-type models do not depend on the order statistics (i.e. the rankings from lowest to highest) of the indices,  $V_{ij}$ . In contrast to this, the probabilities predicted by the exponential choice model depend on both the magnitude of the  $V_{ij}$ s and on the order statistics of each  $V_{ij}$ . Furthermore, not all models with choice probabilities given by a ratio of an exponential term in the numerator divided by a sum of exponential terms in the denominator are logit-type models. Examples of this are the “Random Regret Minimization” and “Relative Advantage Maximization” models where each exponential term depends on variables related to all of the alternatives (Chorus et al., 2014; Leong and Hensher, 2015). In our logit-type models, each exponential term depends only on the attributes of one alternative (through  $S(V_{ij}, \gamma_j)$ ).

While logit-type models generalize many closed-form, individual choice models that have been described in the literature, they are also a particular parametrization of the class of models described by Mattsson et al. (2014). Using the notation of Mattsson et al., we can show equivalence between the two model classes if we set  $w_j = \exp[\tau_j + S(V_j, \gamma_j)]$ , where the index  $i$  has been suppressed to match the notation used in the Mattsson et al. paper (which described the choice probabilities of a single individual). Viewing logit-type models through the lens of the Mattsson et al. paper is useful for two reasons. First, the Mattsson et al. paper provides a rigorous justification for the multinomial specifications of our logit-type models. Secondly, when thinking of further extensions to our work, the Mattsson et al. paper explains why we cannot automatically generalize our logit-type models to models that are analogous to the nested logit model. In particular, Mattsson et al. show that specifying  $S(\cdot)$  is necessary but not sufficient for specifying models that can cope with dependence between one’s random utilities. To account for this dependence (as with nested logit), one



needs to also specify an “aggregation function” that dictates how the various random utilities are combined into a joint distribution. However, determination of such aggregation functions is an open question that we do not attempt to address because it is beyond the scope of this paper.

To recap, the logit-type models introduced in Section 3.1 are both a generalization of many existing models and a special case of a wider class of models introduced by Mattsson et al. (2014). This position allows us to easily extend previously existing binary choice models to the multinomial setting, thereby making them more useful for transportation researchers. Simultaneously, this position allows us to rely on the theoretical justifications that Mattsson et al. provide for our entire class of models. The next subsection will focus on multinomial extensions of binary choice models in greater detail. Looking further ahead, we have noted that there are choice models that are either not part of the logit-type framework or are non-trivial extensions of logit-type models. These non-logit-type models are not considered in this paper, but they point to the need for more research to expand the types of models that avoid the symmetry property described in Section 1. This point will be returned to in Section 6 where we describe the possible future work that stems this paper.

### 3.3. Extending Binary Models to the Multinomial Setting

As noted in the literature review (Section 2), there are many asymmetric, binary choice models, but these models have limited usefulness for transportation researchers because many choice contexts in transportation are inherently multinomial. In this subsection, we propose a technique for using our class of logit-type models given by Equation 2 to create multinomial extensions of existing binary choice models. As a result, transportation scholars and practitioners will be better able to leverage the work that has already been done to create the asymmetric binary choice models that exist in the literature. First, we will describe our procedure, and then we will demonstrate it with two examples. In particular, we will create multinomial generalizations of the binary clog-log model (Yates, 1955) and the binary scobit model (Nagler, 1994). These two models are chosen, in part, because the clog-log model is one of the oldest and most well-known asymmetric discrete choice models (Fisher, 1922; Yates, 1955; McCullagh and Nelder, 1989) and because the scobit model has been used in multiple disciplines such as political science (Nagler, 1994), transportation (Zhang and Timmermans, 2010; Zhang et al., 2011; Wu et al., 2012), and finance (Golet, 2014).

Overall, our procedure for using Equation 2 to extend existing, closed-form, binary choice models to the multinomial setting is given in Table 1. The basic idea behind this procedure is that if we can express an existing, binary probability function as  $\frac{\exp(S_{ij})}{\sum_{\ell=1}^2 \exp(S_{i\ell})}$ , then the work of Mattsson et al. (2014) rigorously shows that there are an infinite number of random utility formulations that could have lead to the given binary choice probabilities. Moreover, Mattsson et al. showed that the same utility formulations, with more alternatives, would lead to choice probabilities of the form given in Equation 2. The extension from a sum of two exponential terms in the denominator of the binary model, to a sum of three or more exponential terms is thereby well-founded.

Following Table 1, we demonstrate our procedure with two examples, deriving multinomial versions of the clog-log and scobit models for the first time. For an example of using this procedure to generalize the binary logit model to the MNL, see Section 3.4.1, where we perform the extension to the multinomial setting in the context of providing a new derivation of the binary logit and the MNL models.

#### 3.3.1. Example 1: Deriving the Multinomial Clog-log Model

The binary clog-log model (Fisher, 1922; Yates, 1955; McCullagh and Nelder, 1989) was introduced within the field of statistics, where there are usually no explanatory variables that vary with one’s alternatives. The probability function of the binary clog-log model is commonly written as  $P_{\text{clog-log}}(y_{ij} = 1 \mid V_{i1}) = 1 - \exp(-e^{V_{i1}})$ , and the function is plotted in both Figures 1 and 2. As statisticians typically do when there are no explanatory variables that vary with one’s alternatives, the probability of the outcome of interest ( $y_{i1} = 1$ ) is spoken of as being only a function of  $V_{i1} = x_{i1}\beta$ , without any regard for  $V_{i2} = x_{i2}\beta$ . Often,  $V_{i2}$  is not even defined. For instance, when statisticians speak of binary logit models, they usually write  $P(y_{i1} = 1 \mid x_{i1}, \beta) = [1 + \exp(-x_{i1}\beta)]^{-1}$  whereas econometricians and transportation researchers would equivalently say  $P(y_{i1} = 1 \mid x_{i1}, x_{i2}, \beta) = [1 + \exp(V_{i2} - V_{i1})]^{-1} = [1 + \exp(\{x_{i2} - x_{i1}\}\beta)]^{-1}$ . Clearly, the unstated assumption is that  $x_{i2} = 0$ . Whenever binary discrete choice models fail to define  $V_{i2}$ , we will adopt the convention that  $V_{i2} = x_{i2}\beta = 0$ . With this in mind, we can express the binary clog-log model as

Table 1: Procedure for Creating Multinomial Extensions of Binary Choice Models

1.	Determine if one's existing binary choice model is given in terms of both $V_{i1}$ and $V_{i2}$ or if it is only given in terms of $V_{i1}$ .
2.	If one's binary choice model is only given in terms of $V_{i1}$ : <ol style="list-style-type: none"> <li>Assume <math>V_{i2} = x_{i2} = 0</math>.</li> <li>Solve for <math>S(V_{i1}, \gamma_1)</math> to identify the functional form of <math>S(\cdot)</math>.</li> <li>Calculate <math>S(V_{i2}, \gamma_2   V_{i2} = 0)</math>.</li> <li>Use <math>S(V_{i1}, \gamma_1)</math> and <math>S(V_{i2}, \gamma_2   V_{i2} = 0)</math> to determine any restrictions on the values of <math>\tau</math> and <math>\gamma</math> that need to be made to establish the binary choice model as a special case of the logit-type models.</li> </ol>
3.	If one's binary choice model is given in terms of both $V_{i1}$ and $V_{i2}$ : <ol style="list-style-type: none"> <li>Express one's existing choice model as a fraction with one term in the numerator and a sum of terms in the denominator.</li> <li>Ensure that each term in the numerator and denominator contains only one index <math>V_{ij}</math>.</li> <li>Directly solve for <math>S_{ij}</math>, for all alternatives <math>j</math>.</li> <li>Determine any restrictions on the values of <math>\tau</math> and <math>\gamma</math> that need to be made to establish the binary choice model as a special case of the logit-type models.</li> </ol>
4.	Relax all restrictions from the previous two steps to generalize the binary model and to create multinomial versions of the model.

a special case of our logit-type models as follows:

$$\begin{aligned}
 P_{\text{clog-log}}(y_{ij} = 1 | x_{i1}, x_{i2} = 0, \beta) &= 1 - \exp(-e^{V_{i1}}) && \text{Step 2a.} \\
 1 - \exp(-e^{V_{i1}}) &\equiv \frac{\exp(S_{i1})}{\sum_{\ell=1}^2 \exp(S_{i\ell})} && \text{Step 2b.} \\
 &= \frac{1}{1 + \exp(S_{i2} - S_{i1})} \\
 &= [1 + \exp(S_{i2} - S_{i1})]^{-1} \\
 [1 - \exp(-e^{V_{i1}})]^{-1} &= 1 + \exp(S_{i2} - S_{i1}) \\
 \left[1 - \frac{1}{\exp(e^{V_{i1}})}\right]^{-1} &= 1 + \exp(S_{i2} - S_{i1}) \\
 \left[\frac{\exp(e^{V_{i1}}) - 1}{\exp(e^{V_{i1}})}\right]^{-1} &= 1 + \exp(S_{i2} - S_{i1}) && (4) \\
 \frac{\exp(e^{V_{i1}})}{\exp(e^{V_{i1}}) - 1} &= 1 + \exp(S_{i2} - S_{i1}) \\
 \frac{\exp(e^{V_{i1}})}{\exp(e^{V_{i1}}) - 1} - \frac{\exp(e^{V_{i1}}) - 1}{\exp(e^{V_{i1}}) - 1} &= \exp(S_{i2} - S_{i1}) \\
 \frac{1}{\exp(e^{V_{i1}}) - 1} &= \exp(S_{i2} - S_{i1}) \\
 [\exp(e^{V_{i1}}) - 1]^{-1} &= \exp(S_{i2} - S_{i1}) \\
 -\ln[\exp(e^{V_{i1}}) - 1] &= S_{i2} - S_{i1} \\
 -\ln[\exp(e^{V_{i1}}) - 1] &= \tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1})
 \end{aligned}$$

On the last line of the right hand side of Equation 4, only  $S(V_{i1})$  involves  $V_{i1}$ . This means that  $S(V_{i1}) = \ln[\exp(e^{V_{i1}}) - 1]$ , and more generally,  $S(V_{ij}) = \ln[\exp(e^{V_{ij}}) - 1]$ . This fact can be derived as follows. First, note that  $S(V_{i1})$  does not contain any arbitrary constants, as these can be thought of as part of  $\tau_1$ .

Next, let  $h(V_{i1}) = \ln [\exp(e^{V_{i1}}) - 1]$ . Then,

$$\begin{aligned}
-h(V_{i1}) &= \tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1}) \\
\frac{\partial(-h(V_{i1}))}{\partial V_{i1}} &= \frac{\partial[\tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1})]}{\partial V_{i1}} \\
-\frac{\partial h(V_{i1})}{\partial V_{i1}} &= -\frac{\partial S(V_{i1})}{\partial V_{i1}} \\
\frac{\partial h(V_{i1})}{\partial V_{i1}} &= \frac{\partial S(V_{i1})}{\partial V_{i1}} \\
\partial h(V_{i1}) &= \partial S(V_{i1}) \\
\int \partial h(V_{i1}) &= \int \partial S(V_{i1}) \\
h(V_{i1}) &= S(V_{i1}) + A && \text{where A is a constant of integration} \\
h(V_{i1}) &= S(V_{i1}) && \text{because } S(V_{i1}) \text{ contains no arbitrary constants} \\
\ln [\exp(e^{V_{i1}}) - 1] &= S(V_{i1})
\end{aligned}$$

With this specification of  $S(\cdot)$ , we can further simplify Equation 4 as follows:

$$\begin{aligned}
-\ln [\exp(e^{V_{i1}}) - 1] &= \tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1}) \\
-S(V_{i1}) &= \tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1}) \\
0 &= \tau_2 + S(V_{i2}) - \tau_1 \\
0 &= \tau_2 + S(0) - \tau_1 && \text{Step 2c.} \\
0 &= \tau_2 + \ln [\exp(e^0) - 1] - \tau_1 \\
0 &= \tau_2 + \ln [e - 1] - \tau_1 \\
\tau_1 - \tau_2 &= \ln [e - 1] && \text{Step 2d.}
\end{aligned} \tag{5}$$

From Equation 5, we have two unknowns  $\tau_1$  and  $\tau_2$ , and one equation. Without loss of generality, we can therefore set  $\tau_2 = 0$  and  $\tau_1 = \ln [e - 1]$ . With these restrictions, we have shown that the binary clog-log model is a special case of the logit type models given by Equation 2, where there are no shape parameters  $\gamma$ , and where  $S(V_{ij}) = \ln [\exp(e^{V_{ij}}) - 1]$ ,  $\tau_1 = \ln [e - 1]$ , and  $\tau_2 = 0$ .

From these results, we can form a “conditional clog-log model” that parallels the “conditional logit model” (McFadden, 1972) and is immediately made useful to transportation researchers and econometricians by allowing explanatory variables that differ across alternatives. To do so, we merely remove the restriction that  $x_{i2} = 0 \forall i$ , and we remove the constraint that  $\tau_1 - \tau_2 = \ln [e - 1]$ . Of course, as with alternative specific constants in general, only the difference  $\tau_2 - \tau_1$  is identified, so one of the two constants should be constrained. This “conditional clog-log model” can easily be extended to the multinomial setting in an analogous fashion to the multinomial logit model. Specifically, the multinomial clog-log model is given by Equation 2, where  $S(V_{ij}, \gamma_j) = S(V_{ij}) = \ln [\exp(e^{V_{ij}}) - 1]$  as derived above, and where as usual, one of the  $\tau_j$ ’s is constrained to zero for identification purposes. For convenience, the probability equation of the multinomial clog-log model is displayed below.

$$P_{\text{clog-log}}(y_{ij} = 1 \mid \tau, V_{i1}, V_{i2}, \dots, V_{ik} \forall \{j, k\} \in C_i) = \frac{\exp(\tau_j + \ln [\exp(e^{V_{ij}}) - 1])}{\sum_{\ell \in C_i} \exp(\tau_\ell + \ln [\exp(e^{V_{i\ell}}) - 1])}$$

### 3.3.2. Example 2: Deriving the Multinomial Scobit Model

The multinomial scobit model is derived from the binary scobit model (see Figure 2) using the same process as with multinomial clog-log model. Given that the binary scobit model (Nagler, 1994) is defined

only in terms of  $V_{i1}$ , we assume  $V_{i2} = x_{i2} = 0$ . From here we write,

$$\begin{aligned}
P_{\text{scobit}}(y_{ij} = 1 \mid x_{i1}, x_{i2} = 0, \beta) &= \frac{1}{(1 + e^{-V_{i1}})^{\gamma_1}}, \quad \gamma_1 \in (0, \infty) && \text{Step 2a.} \\
\frac{1}{(1 + e^{-V_{i1}})^{\gamma_1}} &\equiv \frac{\exp(S_{i1})}{\sum_{\ell=1}^2 \exp(S_{i\ell})} && \text{Step 2b.} \\
&= [1 + \exp(S_{i2} - S_{i1})]^{-1} \\
(1 + e^{-V_{i1}})^{-\gamma_1} &= [1 + \exp(S_{i2} - S_{i1})]^{-1} \\
(1 + e^{-V_{i1}})^{\gamma_1} &= 1 + \exp(S_{i2} - S_{i1}) \\
(1 + e^{-V_{i1}})^{\gamma_1} - 1 &= \exp(S_{i2} - S_{i1}) \\
\ln \left[ (1 + e^{-V_{i1}})^{\gamma_1} - 1 \right] &= S_{i2} - S_{i1} \\
\ln \left[ (1 + e^{-V_{i1}})^{\gamma_1} - 1 \right] &= \tau_2 + S(V_{i2}, \gamma_2) - \tau_1 - S(V_{i1}, \gamma_1), \quad \gamma_2 \in (0, \infty) && (6)
\end{aligned}$$

As before, since  $S(V_{i1}, \gamma_1)$  is the only term on the right hand side of Equation 6 that contains  $V_{i1}$ , we can determine that  $S(V_{i1}, \gamma_1) = -\ln \left[ (1 + e^{-V_{i1}})^{\gamma_1} - 1 \right]$ , and that even more generally,  $S(V_{ij}, \gamma_j) = -\ln \left[ (1 + e^{-V_{ij}})^{\gamma_j} - 1 \right]$ . Substituting these terms back into Equation 6, we can further simplify that equation to:

$$\begin{aligned}
\ln \left[ (1 + e^{-V_{i1}})^{\gamma_1} - 1 \right] &= \tau_2 + S(V_{i2}, \gamma_2) - \tau_1 - S(V_{i1}, \gamma_1) \\
-S(V_{i1}, \gamma_1) &= \tau_2 + S(V_{i2}, \gamma_2) - \tau_1 - S(V_{i1}, \gamma_1) \\
0 &= \tau_2 + S(V_{i2}, \gamma_2) - \tau_1 \\
0 &= \tau_2 - \ln \left[ (1 + e^{-V_{i2}})^{\gamma_2} - 1 \right] - \tau_1 && (7) \\
0 &= \tau_2 - \ln \left[ (1 + e^0)^{\gamma_2} - 1 \right] - \tau_1 && \text{Step 2c.} \\
\tau_1 - \tau_2 &= \ln [2^{\gamma_2} - 1] && \text{Step 2d.}
\end{aligned}$$

Here, we have more unknowns than equations, so some of the parameters are not identified and must be constrained. If we set  $\gamma_2 = 1$ , then this means  $\tau_1 = \tau_2$ , and without loss of generality, we can assume  $\tau_1 = \tau_2 = 0$ . With these constraints, we have shown that the binary scobit model is a special case of the logit-type models given by Equation 2.

As with the binary clog-log model, the binary scobit model can be immediately generalized to a “conditional scobit model” that allows for explanatory variables that differ across alternatives. The “conditional scobit model” is derived by removing the constraints  $\gamma_2 = 1$ ,  $x_{i2} = 0$ , and  $\tau_1 = \tau_2 = 0$ . As usual, one of the alternative specific constants,  $\tau_1$  or  $\tau_2$ , must still be constrained for identification purposes.

Finally, as with the multinomial clog-log model, the generalization of the “conditional scobit model” to the multinomial setting is immediate. The multinomial scobit model is given by Equation 2, where  $S(V_{ij}, \gamma_j) = -\ln \left[ (1 + e^{-V_{ij}})^{\gamma_j} - 1 \right]$  and where  $\gamma_j$  is a scalar, for each alternative  $j$ , that is to be estimated along with  $\beta$  and all but one of the  $\tau_j$ ’s (for identifiability). For convenience, the probability formula for the multinomial scobit model is displayed below.

$$P_{\text{scobit}}(y_{ij} = 1 \mid \tau, \gamma, V_{i1}, V_{i2}, \dots, V_{ik} \forall \{j, k\} \in C_i) = \frac{\exp(\tau_j - \ln \left[ (1 + e^{-V_{ij}})^{\gamma_j} - 1 \right])}{\sum_{\ell \in C_i} \exp(\tau_\ell - \ln \left[ (1 + e^{-V_{i\ell}})^{\gamma_\ell} - 1 \right])}$$

### 3.4. Creating New Asymmetric Choice Models

In Section 3.3, we showed how one can extend existing, binary choice models to the multinomial setting. In this section, we will present our method for creating new binary choice models. Note that our proposed process is general enough to create both new asymmetric and new symmetric choice models. Together, Section 3.3 and Section 3.4 provide a way to create new multinomial choice models. In this paper, however,

we will focus on the creation of new asymmetric, multinomial choice models. The rest of this subsection will proceed as follows. First, we will briefly review traditional methods in transportation for creating new choice models, and why we think such methods are not easy to use. Next, we will present an alternative approach for creating new binary choice models. We will then review the key concepts necessary to understand this approach, and finally, we will present two examples where we demonstrate the procedure by creating new, asymmetric, binary choice models and extending them to the multinomial setting.

As noted by Ben-Akiva and Lerman, “varying the assumptions about the distributions of [one’s utilities] [...] leads to different choice models” (Ben-Akiva and Lerman, 1985, p.65). This approach of first specifying the distribution of one’s utilities, and then deriving one’s choice probabilities, is commonly used in transportation. For instance, it is used by the transportation researchers cited above such as Castillo et al. (2008), Fosgerau and Bierlaire (2009), Li (2011), and Mattsson et al. (2014). While clearly a viable approach, discrete choice analysts have acknowledged that “it will often be difficult to make strong statements about the overall distribution of [one’s utilities]” (Ben-Akiva and Lerman, 1985, p.66). To sidestep these difficulties Daniel McFadden (emphasis is his own) wrote that:

“In practice, it is difficult to define joint distributions [of one’s utilities] which allow the computation of econometrically useful formulas for the [selection probabilities]. An alternative approach is to specify formulas for the selection probabilities and then examine the question of whether these formulas could be obtained [...] from *some* distribution of utility-maximizing consumers” (McFadden, 1972, p.108).

This approach of directly specifying probability formulas is the one that we will take. From the work of Mattsson et al. (2014), we know that any choice model of the form given in Equation 2 can be generated from an infinite number of joint distributions of one’s utilities. Moreover, we know that the newly derived logit-type models will be “well-behaved.” To be specific, because logit type models have an exponential term for their numerator and a sum of exponential terms for their denominator, where the sum includes the numerator, logit-type models will always return probabilities between zero and one. Also, because  $x_{ij}$  only appears in  $S_{ij}$  and because  $S_{ij}$  was defined as being a monotonically increasing function of  $V_{ij} = x_{ij}\beta$ , interpreting whether the probability of choosing alternative  $j$  increases or decreases when we increase a variable in  $x_{ij}$  remains as easy as it was with the standard MNL model. Often, such interpretation consists of just knowing the sign on the index coefficient of the variable of interest. Given these beneficial properties, the question that now remains, for the purposes of generating new logit-type models, is “how should one specify  $S(\cdot)$ ?”

To specify the  $S(\cdot)$  function in one’s logit-type models, we created the three-step procedure<sup>6</sup> shown in Table 2. Note that this procedure will make use of potentially unfamiliar terms and concepts such as “binary loss functions,” “asymmetric loss functions,” “properties of loss functions,” and “related, binary probability functions.” However, all of these terms will be explained and made more precise in the following paragraphs. After these explanations, we will demonstrate our procedure. First, we will use the process in Table 2 to re-derive the familiar MNL model. Then we will further demonstrate the procedure by creating two new, closed-form, asymmetric probability functions.

Given that the first step in our proposed procedure is to choose a binary loss function with properties that are desirable for one’s study, we will begin by defining loss functions, and then we will explain what is meant by properties of the loss function. Loss functions are functions that measure the quality of one’s predictions, and binary loss functions measure the quality of one’s predictions when one’s observed, dependent variable takes on one of two possible values. Overall, there are two types of binary loss functions: “class probability estimation (CPE) loss functions” and “composite loss functions” (Reid and Williamson, 2010). For the purposes of Step 1 of our procedure, either of the two types of loss functions may be chosen. However, the two types of loss functions lead to differences in how the related probability functions are derived in Step 2 of our procedure. As a result, we will briefly describe both types of losses in the next paragraph. Additionally,

---

<sup>6</sup>Note that our procedure was motivated by computer scientists who made use of asymmetric loss functions (defined in the coming paragraphs) when dealing with class-imbalanced datasets. When investigating this use of asymmetric loss functions, we came across literature that noted the fact that loss functions are related to specific probability functions. This discovery lead us to think that a useful way of deriving choice models, given the literature on choosing or designing loss functions, would be to first choose a desired loss function and then derive its related probability function.

Table 2: Procedure for Creating New Multinomial Choice Models

1.	Choose a <i>binary loss function</i> with properties that are desirable for one’s study. If an asymmetric choice model is desired, then be sure to choose an <i>asymmetric loss function</i> .
2.	Derive the <i>related, binary probability function</i> for one’s chosen loss function.
3.	Use the procedure detailed in Section 3.3 to convert one’s derived probability function to a logit-type model. In the process, one will have determined $S(\cdot)$ and created a new multinomial choice model.

we will make connections with concepts that most readers will be familiar with by showing the CPE and composite loss functions that are related to the binary logit model.

We will start with CPE loss functions. CPE losses take an observed outcome and the predicted probability of that outcome occurring as arguments, and they output a penalty (i.e. a non-negative value) for discrepancies between the observation and prediction (Reid and Williamson, 2010). Typically, the returned penalty increases as the magnitude of the discrepancy increases. For example, the CPE loss function that is related to the binary logit model is the negative log-likelihood. This CPE loss is given by

$$\text{Negative Log-Likelihood}(y_{i1}, P(y_{i1} = 1 \mid V_{i1}, V_{i2})) = 1_{\{y_{i1}=1\}} (-\ln[P(y_{i1} = 1 \mid V_{i1}, V_{i2})]) + 1_{\{y_{i1}=0\}} (-\ln[1 - P(y_{i1} = 1 \mid V_{i1}, V_{i2})]) \quad (8)$$

where  $1_{\{r\}}$  is an indicator function that equals 1 if  $r$  is true and 0 otherwise.

Moving to composite losses, we noted in Section 3.3.1 that statisticians and computer scientists often speak of  $P(y_{i1} = 1)$  as being only a function of  $V_{i1} = x_{i1}\beta$ , without any regard for  $V_{i2} = x_{i2}\beta$ . In such settings, where it is often implicitly the case that  $x_{i2} = 0$ , one can speak of “composite loss functions,” that are simply functions of  $V_{i1}$ . Formally, composite loss functions are CPE loss functions composed of the probability function  $P(y_{i1} = 1 \mid V_{i1})$  (Reid and Williamson, 2010). As an example, consider the composite loss function that is related to the binary logit model—the log-loss. This loss function is derived by composing the negative log-likelihood given in Equation 8 with the probability function,  $P(y_{i1} = 1 \mid V_{i1}) = [1 + \exp(-V_{i1})]^{-1}$ . We will omit the algebra used to simplify the composition, but the log-loss is given by the following formula:

$$\text{Log-Loss}(y_{i1}, V_{i1}) = 1_{\{y_{i1}=1\}} \ln(1 + e^{-V_{i1}}) + 1_{\{y_{i1}=0\}} \ln(1 + e^{V_{i1}}) \quad (9)$$

Given the formulation of composite losses, these functions differ from CPE loss functions only in their arguments. While both losses return a penalty for the discrepancy between one’s observed outcome and the predicted probability of that outcome occurring, composite loss functions take the observed outcome and  $V_{i1}$  (as opposed to  $P(y_{i1} = 1 \mid V_{i1})$ ) as arguments. Note that CPE loss functions are defined for arbitrary probability functions, including those of the form  $P(y_{i1} = 1 \mid V_{i1}, V_{i2})$ , whereas composite loss functions are only defined for probability functions of the form  $P(y_{i1} = 1 \mid V_{i1})$ . This is analogous to the situation described in Section 3.3 where one’s probability function could depend only on  $V_{i1}$  or on both  $V_{i1}$  and  $V_{i2}$ . As in Table 1, different steps are taken based on the situation we are in.

Now, beyond merely choosing a loss function, step 1 of our procedure requires choosing a loss function based on its properties. To place such properties in context, we emphasize that for our purposes, the most important use of loss functions is as a tool for parameter estimation. In an optimization setting, loss functions are used in statistics and computer science to estimate parameters of interest, such as the  $\beta$ ’s in one’s choice model (Gneiting and Raftery, 2007; Dawid, 2006). The idea is that one chooses the set of parameters that minimizes the total loss (i.e. the sum of the loss for each observation), given one’s dataset. In a parameter estimation setting, each loss function has properties that impact the estimation process and results. One important property is whether or not a loss function is symmetric. Symmetric, binary loss functions output equal-magnitude penalties for equal magnitude discrepancies, regardless of the observed outcome. For example, imagine we are analyzing the losses incurred on two observations: observations 1 and 2. Observation 1 is associated with outcome 1, and observation 2 is associated with outcome 2. For both observations, we predicted a 30% probability of that observation being associated with its actual outcome. A symmetric loss function would assign the same penalty to our predictions for both observation 1 and

observation 2. In contrast, an asymmetric loss function would assign different penalties to observation 1 and observation 2 because asymmetric loss functions treat discrepancies unequally across the two possible outcomes.

Aside from symmetry, loss functions have other properties that impact one’s parameter estimates. For example, one might consider whether one’s loss function is strictly proper (i.e. the loss is Fisher consistent and increasing discrepancies *always* lead to increasing penalties) (Buja et al., 2005; Reid and Williamson, 2010); robust against outliers (Pregibon, 1982; Carroll and Pederson, 1993; Bianco and Yohai, 1996); or sparsity-inducing (in terms of identifying “irrelevant” predictors”, i.e. setting their  $\beta$  coefficient to zero) (Kyung et al., 2010; Bach et al., 2012; Xu et al., 2012). In general, there are numerous properties that might be of interest. As such, it is beyond the scope of this paper to (1) comprehensively review and describe these properties or (2) instruct readers on how to design their loss functions with respect to these various properties. Interested readers seeking guidance may refer to works such as Hennig and Kutlukaya (2007) or Merkle and Steyvers (2013). Our main point is that loss functions have properties, that analysts can choose the most desirable mix of properties for their research needs, and that once an analyst has designed or found a loss function with the appropriate properties for their study, a related probability function can be derived from the chosen loss function. The next paragraph will describe precisely what is meant by the term “related probability function” and how one can derive it.

In general, one can derive unique probability functions from both composite loss functions and strictly proper CPE loss functions. In the case of CPE loss functions, the related probability function is such that when minimizing one’s total loss, one is guaranteed to have an optimization problem that is convex in one’s  $\beta$ ’s (Buja et al., 2005; Reid and Williamson, 2010). In the case of composite loss functions, the related probability function is the one that must have been used to derive the composite loss (Reid and Williamson, 2010). In each case, the derivation of the related probability functions uses what are known as the partial losses for a binary loss function. The partial losses are simply the functions used to supply the penalty for predictions on each of the two possible discrete outcomes (Buja et al., 2005; Reid and Williamson, 2010). Formally, a given binary loss function  $L(y_{i1}, \cdot)$  can be written as  $L(y_{i1}, \cdot) = 1_{\{y_{i1}=1\}}L_1(\cdot) + 1_{\{y_{i1}=0\}}L_2(\cdot)$ . The second argument of  $L(y_{i1}, \cdot)$  depends on whether or not we are using a CPE loss function or a composite loss function. In either case,  $L_1(\cdot)$  and  $L_2(\cdot)$  are known as the partial losses for  $L(y_{i1}, \cdot)$ .  $L_1$  determines the penalty if the observation is associated with outcome 1, and  $L_2$  determines the penalty if the observation is associated with outcome 2. To derive the related probability functions we will start with the simpler derivation, the one for composite loss functions. It has been proven that in order for a binary composite loss function with differentiable partial losses to have been created by the composition of a CPE loss function and a probability function, the probability function must satisfy the following criteria (Reid and Williamson, 2010, Eq. 11):

$$P(y_{i1} = 1 | V_{i1}) = \frac{L'_2(V_{i1})}{L'_2(V_{i1}) - L'_1(V_{i1})} \quad (10)$$

We will use this equation directly in order to derive the related probability function for composite losses. For strictly proper CPE loss functions, if one makes the usual assumption from statistics and computer science that  $V_{i2} = 0$ , then there exists a canonical probability function that can be derived by solving the following differential equation<sup>7</sup> for  $P(y_{i1} = 1 | V_{i1}, V_{i2} = 0)$ :

$$\frac{d[L_2(\hat{p}(V_{i1}))]}{d\hat{p}(V_{i1})} = \frac{\hat{p}(V_{i1})}{\hat{p}'(V_{i1})} \quad \text{where } \hat{p}(V_{i1}) = P(y_{i1} = 1 | V_{i1}, V_{i2} = 0) \quad (11)$$

In the following examples, we will show how both of these equations can be used in our proposed procedure for generating new multinomial choice models. Our first example will derive the MNL model using both its related CPE loss (the negative log-likelihood) and its related composite loss (the log-loss). Following those derivations, we will use our proposed procedure to create two new, asymmetric, closed-form probability functions. In particular, we will create the *uneven logit model* from a composite loss function and then create the *asymmetric logit model* using a CPE loss function.

---

<sup>7</sup>Our derivation of this formula is given in the appendix.

### 3.4.1. Example 3: Deriving the MNL Model

In this subsection, we aim to clarify the procedures in Table 2 by deriving the familiar MNL model using both its related CPE loss and its related composite loss. We start with the composite loss, since it is a more straightforward derivation.

#### Deriving the Binary Logit Model Using the Log-Loss

In Step 1, we are required to choose a binary loss function. For the binary logit model, one such loss function is the log-loss (i.e. the related composite loss for the binary logit model). As shown in Equation 9, the binary log-loss is:

$$\begin{aligned}\text{Log-Loss}(y_{i1}, V_{i1}) &= 1_{\{y_{i1}=1\}} \ln(1 + e^{-V_{i1}}) + 1_{\{y_{i1}=0\}} \ln(1 + e^{V_{i1}}) \\ &= 1_{\{y_{i1}=1\}} L_1(V_{i1}) + 1_{\{y_{i1}=0\}} L_2(V_{i1})\end{aligned}$$

The necessary derivatives for Step 2 are  $L'_1(V_{i1})$  and  $L'_2(V_{i1})$ . For the log-loss, these derivatives are:

$$\begin{aligned}L'_1(V_{i1}) &= \frac{-e^{-V_{i1}}}{1 + e^{-V_{i1}}} \\ &= \frac{-1}{1 + e^{V_{i1}}} \\ L'_2(V_{i1}) &= \frac{e^{V_{i1}}}{1 + e^{V_{i1}}}\end{aligned}$$

Below, we use these derivatives to derive the formula for binary logit model that is commonly used in statistics and computer science applications (where  $V_{i2}$  implicitly equals zero):

$$\begin{aligned}P_{\text{binary logit}}(y_{i1} = 1 \mid V_{i1}) &= \frac{L'_2(V_{i1})}{L'_2(V_{i1}) - L'_1(V_{i1})} \\ &= \frac{e^{V_{i1}}}{\frac{1 + e^{V_{i1}}}{e^{V_{i1}}} - \frac{-1}{1 + e^{V_{i1}}}} \\ &= \frac{1 + e^{V_{i1}}}{\frac{e^{V_{i1}} + 1}{1 + e^{V_{i1}}}} \\ &= \frac{e^{V_{i1}}}{1 + e^{V_{i1}}}\end{aligned} \tag{12}$$

For a moment, we will defer Step 3 (where we extend the binary logit model to the multinomial setting). Instead we will now show how the same binary logit model formula can be obtained from the negative log-likelihood. The final step of extending the binary logit model to the MNL model will be the same for both versions of the procedure in Table 2, regardless of whether we start with a CPE loss or a composite loss function.

#### Deriving the Binary Logit Model Using the Negative Log-Likelihood

Similar to the use of the log-loss, we can use the negative log-likelihood as our CPE loss function from which we derive the binary logit model. As given in Equation 8, the negative log-likelihood with  $V_{i2}$  assumed to equal zero is

$$\begin{aligned}\text{Negative Log-Likelihood}(y_{i1}, P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)) &= 1_{\{y_{i1}=1\}} (-\ln[P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)]) + \\ &\quad 1_{\{y_{i1}=0\}} (-\ln[1 - P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)]) \\ &= 1_{\{y_{i1}=1\}} L_1[P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)] + \\ &\quad 1_{\{y_{i1}=0\}} L_2[P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)]\end{aligned}$$



For Step 2, we need the derivative of  $L_2$  with respect to  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)$ . This derivative is:

$$\frac{\partial L_2 [P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)]}{\partial P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)} = \frac{1}{1 - P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)}$$

From here, we can use Equation 11 to solve for  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)$  as follows. Let  $\hat{p}(V_{i1}) = P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)$ . Then,

$$\begin{aligned} \frac{\partial L_2 [\hat{p}(V_{i1})]}{\partial \hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\frac{\partial \hat{p}(V_{i1})}{\partial V_{i1}}} \\ \frac{1}{1 - \hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\frac{\partial \hat{p}(V_{i1})}{\partial V_{i1}}} \\ \frac{1}{\hat{p}(V_{i1}) [1 - \hat{p}(V_{i1})]} &= \frac{\partial V_{i1}}{\partial \hat{p}(V_{i1})} \\ \frac{\partial \hat{p}(V_{i1})}{\hat{p}(V_{i1}) [1 - \hat{p}(V_{i1})]} &= \partial V_{i1} \\ \left[ \frac{1}{\hat{p}(V_{i1})} + \frac{1}{1 - \hat{p}(V_{i1})} \right] \partial \hat{p}(V_{i1}) &= \partial V_{i1} \\ \int \left[ \frac{1}{\hat{p}(V_{i1})} + \frac{1}{1 - \hat{p}(V_{i1})} \right] \partial \hat{p}(V_{i1}) &= \int \partial V_{i1} \\ \ln [\hat{p}(V_{i1})] - \ln [1 - \hat{p}(V_{i1})] &= V_{i1} + A \quad \text{where } A \text{ is a constant of integration} \\ \ln \left[ \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right] &= V_{i1} + A \end{aligned}$$

As with any differential equation, we need a boundary condition to be able to determine the value of  $A$ . A typical condition would be that  $\hat{p}(V_{i1} = 0) = 0.5$ . With this boundary condition,  $A = 0$  and we have

$$\ln \left[ \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right] = V_{i1}$$

Standard algebraic manipulation leads back to Equation 12 for the binary logit model where  $V_{i2}$  is assumed to be zero.

#### *Extending Binary Logit to MNL*

Finally, Step 3 of our procedure for creating new choice models is that we use the procedure from Section 3.3 to create a multinomial extension of the binary version of our model. This is done below. The labels to the right of the equations refer to the steps in Table 1.

$$\begin{aligned} P_{\text{binary logit}}(y_{ij} = 1 \mid x_{i1}, x_{i2} = 0, \beta) &= \frac{e^{V_{i1}}}{1 + e^{V_{i1}}} && \text{Step 2a.} \\ \frac{e^{V_{i1}}}{1 + e^{V_{i1}}} &\equiv \frac{\exp(S_{i1})}{\sum_{\ell=1}^2 \exp(S_{i\ell})} && \text{Step 2b.} \\ \frac{1}{1 + e^{-V_{i1}}} &= \frac{1}{1 + e^{S_{i2} - S_{i1}}} && (13) \\ 1 + e^{-V_{i1}} &= 1 + e^{S_{i2} - S_{i1}} \\ -V_{i1} &= S_{i2} - S_{i1} \\ -V_{i1} &= \tau_2 + S(V_{i2}) - \tau_1 - S(V_{i1}) \end{aligned}$$

Using the same arguments as in Section 3.3, we find  $S(V_{ij}) = V_{ij}$ . Substituting this equality back into the last line of Equation 13, we get:

$$\begin{aligned}
-V_{i1} &= \tau_2 + V_{i2} - \tau_1 - V_{i1} \\
0 &= \tau_2 + V_{i2} - \tau_1 \\
0 &= \tau_2 - \tau_1 \quad \text{because } V_{i2} = 0 && \text{Step 2c.} \\
\tau_1 &= \tau_2 && \text{Step 2d.}
\end{aligned} \tag{14}$$

The two constants,  $\tau_1$  and  $\tau_2$  are not identified. Without loss of generality we can set  $\tau_1$ , and implicitly  $\tau_2$ , equal to zero. This establishes the binary logit model as a special case of our logit-type models. The generalization to the MNL model given in Equation 1 follows by removing the restrictions on  $\tau$  and using Equation 2 with  $S(V_{ij}) = V_{ij}, \forall j \in C_i$ , subject to identification. Typically, researchers include an alternative specific constant in  $x_{ij}$ . Such a constant will cause a lack of identification with  $\tau$  in the MNL model. In such conditions, one can set  $\tau_j = 0 \forall j$ , and the MNL formula from Equation 1 is recovered exactly.

Now that the process of deriving a multinomial choice model from a loss function has been reviewed for the MNL model, we will further illustrate this process by creating two new, multinomial, asymmetric choice models. The creation of these two new models is shown in the following two subsections, Section 3.4.2 and Section 3.4.3.

#### 3.4.2. Example 4: Creating the Uneven Logit Model

In ‘‘Calibrated asymmetric surrogate losses’’ (Scott, 2012), Scott provides a way of creating asymmetric, composite loss functions from symmetric ones. Scott’s main goal was to create loss functions that performed optimally under different costs for wrong classification predictions (i.e. binary predictions as opposed to probability predictions). Since altering misclassification costs is one technique used to deal with class imbalance in computer science, we decided to see whether the probability functions derived from Scott’s asymmetric composite losses would be useful for making probability predictions under class imbalance.

To begin, we applied the procedures in Scott’s paper to the log-loss to derive the following ‘‘uneven log-loss’’:

$$\text{Uneven log-loss} = 1_{\{y_{i1}=1\}} \ln(1 + e^{-V_{i1}}) + 1_{\{y_{i1}=0\}} \frac{1}{\gamma_1} \ln(1 + e^{\gamma_1 V_{i1}}), \quad \gamma_1 > 0 \tag{15}$$

Here,  $L_1(V_{i1}) = \ln(1 + e^{-V_{i1}})$  and  $L_2(V_{i1}) = (\gamma_1)^{-1} \ln(1 + e^{\gamma_1 V_{i1}})$ . Using these equations, and the derivatives of the partial losses, we derive the related probability function as follows:

$$\begin{aligned}
L'_1(V_{i1}) &= \frac{-e^{-V_{i1}}}{1 + e^{-V_{i1}}} \\
L'_2(V_{i1}) &= \frac{1}{1 + e^{-\gamma_1 V_{i1}}} \\
P(y_{i1} = 1 \mid V_{i1}) &= \frac{L'_2(V_{i1})}{L'_2(V_{i1}) - L'_1(V_{i1})} \\
&= \frac{1}{1 - \frac{L'_1(V_{i1})}{L'_2(V_{i1})}} \\
&= \frac{1}{1 + \left( \frac{1 + e^{-\gamma_1 V_{i1}}}{1 + e^{-V_{i1}}} \right) e^{-V_{i1}}}
\end{aligned} \tag{16}$$

Because we derived this probability function from the uneven log-loss, we named it the ‘‘uneven logit model.’’ To visualize the range of possible shapes that the binary, uneven logit model can take, see Figure 2.

Now, using the procedure from Table 1, we can convert the probability function derived in Equation 16 into a logit-type model as given in Equation 2. We will omit the algebra, but the result is that we find  $S(V_{ij}, \gamma_j) = V_{ij} + \ln(1 + e^{-V_{ij}}) - \ln(1 + e^{-\gamma_j V_{ij}})$  and  $\tau_j = 0 \forall j$ . Note  $\gamma_j$ , for all alternatives  $j$ , is still required to be positive because this ensures that  $S_{ij}$  is monotonically increasing in  $V_{ij}$ .

As with the multinomial clog-log and multinomial scobit models, we can immediately generalize the uneven logit model to a conditional uneven logit model. This is done simply by allowing  $x_{i2} \neq 0$  and  $\tau_j \neq 0$ ,

although one of the  $\tau_j$ 's must still be constrained for identification purposes. Lastly, the multinomial uneven logit model is immediately obtained by using Equation 2 with  $S(\cdot)$  as derived in the last paragraph. As with the multinomial clog-log and scobit models, the probability function for the multinomial uneven logit model is displayed below for convenience.

$$P_{\text{uneven logit}}(y_{ij} = 1 \mid \tau, \gamma, V_{i1}, V_{i2}, \dots, V_{ik} \forall \{j, k\} \in C_i) = \frac{\exp[\tau_j + V_{ij} + \ln(1 + e^{-V_{ij}}) - \ln(1 + e^{-\gamma_j V_{ij}})]}{\sum_{\ell \in C_i} \exp[\tau_\ell + V_{i\ell} + \ln(1 + e^{-V_{i\ell}}) - \ln(1 + e^{-\gamma_\ell V_{i\ell}})]}$$

### 3.4.3. Example 5: Creating the Asymmetric Logit Model

Similar to Scott (2012), Winkler (1994) in his paper “Evaluating Probabilities: Asymmetric Scoring Rules” developed a methodology for creating asymmetric loss functions from symmetric loss functions<sup>8</sup>. However, unlike Scott, Winkler wanted to account for differing states of knowledge as opposed to different misclassification costs. In particular, Winkler wanted a loss function whose risk<sup>9</sup> was maximized at the probability that corresponds to “knowing nothing” (Winkler, 1994). This would allow one to judge probability forecasts in a way that accounts for the fact that knowing “nothing” does not always mean assigning a 50% probability to the outcome of interest. Sometimes an analyst may still know that (on average) individuals have a greater or lesser than 50% chance of choosing a given alternative. One such case where this is true is in class imbalanced situations. Given the link between Winkler’s motivation for developing his asymmetric scoring rules and the class imbalanced scenarios that motivated this paper, we decided to investigate whether the probability functions derived from Winkler’s asymmetric losses would be useful for making probability predictions under class imbalance.

Applying Winkler’s methods to the negative log-likelihood, and making the assumption that  $V_{i2} = 0$ , leads to the following asymmetric, negative log-likelihood:

$$\begin{aligned} \text{Asymmetric, Negative Log-Likelihood} &= 1_{\{y_{i1}=1\}} L_1(P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)) + \\ &\quad 1_{\{y_{i1}=0\}} L_2(P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)) \\ L_1(P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)) &= \begin{cases} \frac{\ln(\gamma_1) - \ln[P(y_{i1}=1 \mid V_{i1}, V_{i2}=0)]}{-\ln(\gamma_1)}, & P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) \geq \gamma_1 \\ \frac{\ln(\gamma_1) - \ln[P(y_{i1}=1 \mid V_{i1}, V_{i2}=0)]}{-\ln(1-\gamma_1)}, & P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) < \gamma_1 \end{cases} \\ L_2(P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)) &= \begin{cases} \frac{\ln(1-\gamma_1) - \ln[1-P(y_{i1}=1 \mid V_{i1}, V_{i2}=0)]}{-\ln(\gamma_1)}, & P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) \geq \gamma_1 \\ \frac{\ln(1-\gamma_1) - \ln[1-P(y_{i1}=1 \mid V_{i1}, V_{i2}=0)]}{-\ln(1-\gamma_1)}, & P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) < \gamma_1 \end{cases} \\ &\quad \text{where } \gamma_1 \in (0, 1) \end{aligned} \tag{17}$$

Because the asymmetric, negative log-likelihood is piecewise defined, deriving the related probability function requires us to solve Equation 11 twice, once for each case:  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)$  greater than or equal to  $\gamma_1$ , and  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0)$  less than  $\gamma_1$ . The result will be a piecewise defined probability function. However, we need to avoid circular reasoning when defining the pieces of the probability function. In particular, we cannot define the pieces of the probability function using conditions based on the value of the probability function, as is done in the asymmetric, negative log-likelihood. To construct conditions for the related probability function, we note that Equation 11 is a differential equation, so we will need boundary conditions to identify the constant of integration. Our boundary condition for the two cases will be that  $P(y_{i1} = 1 \mid V_{i1} = 0, V_{i2} = 0)$  equals  $\gamma_1$ . This condition will ensure continuity of the resulting probability function. Moreover, when combined with the fact that the probability function is monotonically increasing in  $V_{i1}$ , this boundary condition allows us to express the pieces of the probability function in terms of  $V_{i1} \geq 0$  (which implies  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) \geq \gamma_1$ ) and  $V_{i1} < 0$  (which implies  $P(y_{i1} = 1 \mid V_{i1}, V_{i2} = 0) < \gamma_1$ ).

<sup>8</sup>Technically, Winkler developed a method for constructing asymmetric scoring rules from symmetric scoring rules. However, scoring rules are simply negated loss functions, so Winkler’s methods also allow one to create asymmetric loss functions.

<sup>9</sup>Note the risk of a loss function is the expectation of the loss over all possible datasets, given the true parameters being estimated (Keener, 2006).

Starting with the case,  $P(y_{i1} = 1 | V_{i1}, V_{i2} = 0) \geq \gamma_1$  and  $P(y_{i1} = 1 | V_{i1} = 0, V_{i2} = 0) = \gamma_1$ , we have:

$$\begin{aligned}
\frac{d[L_2(\hat{p}(V_{i1}))]}{d\hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\hat{p}'(V_{i1})} \quad \text{where } \hat{p}(V_{i1}) = P(y_{i1} = 1 | V_{i1}, V_{i2} = 0) \\
\left[ \frac{-1}{\ln(\gamma_1)} \right] \frac{1}{1 - \hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\hat{p}'(V_{i1})} \\
\left[ \frac{-1}{\ln(\gamma_1)} \right] \int \frac{d\hat{p}}{\hat{p}(V_{i1}) [1 - \hat{p}(V_{i1})]} &= \int dv \\
\left[ \frac{-1}{\ln(\gamma_1)} \right] \ln \left( \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right) &= v + A \quad \text{where } A \text{ is a constant} \\
\left[ \frac{-1}{\ln(\gamma_1)} \right] \ln \left( \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right) &= v - \left[ \frac{1}{\ln(\gamma_1)} \right] \ln \left( \frac{\gamma_1}{1 - \gamma_1} \right) \\
\text{which simplifies to } \hat{p}(V_{i1}) &= \frac{1}{1 + (\gamma_1^{-1} - 1) \gamma_1^{V_{i1}}}, \quad V_{i1} \geq 0, \gamma_1 \in (0, 1)
\end{aligned} \tag{18}$$

Similarly, for the case where  $P(y_{i1} = 1 | V_{i1}, V_{i2} = 0) < \gamma_1$  and  $P(y_{i1} = 1 | V_{i1} = 0, V_{i2} = 0) = \gamma_1$ , we have:

$$\begin{aligned}
\frac{d[L_2(\hat{p}(V_{i1}))]}{d\hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\hat{p}'(V_{i1})} \quad \text{where } \hat{p}(V_{i1}) = P(y_{i1} = 1 | V_{i1}, V_{i2} = 0) \\
\left[ \frac{-1}{\ln(1 - \gamma_1)} \right] \frac{1}{1 - \hat{p}(V_{i1})} &= \frac{\hat{p}(V_{i1})}{\hat{p}'(V_{i1})} \\
\left[ \frac{-1}{\ln(1 - \gamma_1)} \right] \int \frac{d\hat{p}}{\hat{p}(V_{i1}) [1 - \hat{p}(V_{i1})]} &= \int dv \\
\left[ \frac{-1}{\ln(1 - \gamma_1)} \right] \ln \left( \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right) &= v + B \quad \text{where } B \text{ is a constant} \\
\left[ \frac{-1}{\ln(1 - \gamma_1)} \right] \ln \left( \frac{\hat{p}(V_{i1})}{1 - \hat{p}(V_{i1})} \right) &= v - \left[ \frac{1}{\ln(1 - \gamma_1)} \right] \ln \left( \frac{\gamma_1}{1 - \gamma_1} \right) \\
\text{which simplifies to } \hat{p}(V_{i1}) &= \frac{1}{1 + \gamma_1^{-1} (1 - \gamma_1)^{V_{i1} + 1}}, \quad V_{i1} < 0, \gamma_1 \in (0, 1)
\end{aligned} \tag{19}$$

Together, the binary asymmetric logit model can be written as:

$$P(y_{ij} = 1 | V_{i1}, V_{i2} = 0) = \begin{cases} \frac{1}{1 + (\gamma_1^{-1} - 1) \gamma_1^{V_{i1}}}, & V_{i1} \geq 0 \\ \frac{1}{1 + \gamma_1^{-1} (1 - \gamma_1)^{V_{i1} + 1}}, & V_{i1} < 0 \end{cases} \tag{20}$$

For the readers' convenience, the binary asymmetric logit model is displayed in Figure 2, where we aim to highlight the range of possible shapes that the model can take. Note, we named this probability function the "asymmetric logit model" because it is derived from the asymmetric, negative log-likelihood.

Now, following the procedure in Table 1, we can show that for the binary asymmetric logit model,

$$\begin{aligned}
S(V_{ij}, \gamma_j) &= \begin{cases} \ln(\gamma_j) - V_{ij} \ln(\gamma_j) & V_{ij} \geq 0 \\ \ln(\gamma_j) - V_{ij} \ln(1 - \gamma_j) & V_{ij} < 0 \end{cases} \\
\tau_j &= 0, \quad \forall j
\end{aligned}$$

With these expressions, we can proceed from the binary case to the conditional and multinomial cases. In doing so, however, we must take care to generalize the restrictions and boundary condition used to derive the binary asymmetric logit model. In particular, we will require that

- $\gamma_j \in (0, 1) \quad \forall j$ ,
- $\sum_j \gamma_j = 1$ ,

- $P(y_{ij} = 1 \mid V_{ik} = 0, \forall k \in C_i) = \gamma_j \forall j$ , and
- that the multinomial, asymmetric logit model nest the multinomial logit model the same way the binary, asymmetric logit model nests the binary logit model<sup>10</sup>.

With all of these requirements, the multinomial, asymmetric logit model can be written as given in Equation 2, where

$$S(V_{ij}, \gamma_j) = \begin{cases} \ln(\gamma_j) - V_{ij} \ln(\gamma_j), & V_{ij} \geq 0 \\ \ln(\gamma_j) - V_{ij} \ln\left(\frac{1 - \gamma_j}{J - 1}\right), & V_{ij} < 0 \end{cases} \quad (21)$$

where  $J$  = The total number of possible alternatives in one's dataset

and where one of the  $\gamma_j$ 's must be constrained for identification purposes. Note that unlike the multinomial clog-log, scobit, and uneven logit models, we will not display the probability function for the multinomial asymmetric logit model. Because of the piecewise definition of each  $S_{ij}$ , there are  $2^J$  probability functions where  $J$  is the total number of possible alternatives in the dataset. In other words, there is one function for each of the possible permutations of the indices ( $V_{ij}$ ) being positive or negative. Thus, even for three alternatives, we would need to display 8 equations. The simplest way of stating the multinomial asymmetric logit model is to refer to Equation 2 and note that  $S_{ij}$  is piecewise defined for all  $j$  in this model.

### 3.5. Summary

To summarize, Section 3.1 presented our proposed class of logit-type models and showed how they avoid the symmetry property described in the introduction. Section 3.2 then positioned our logit-type models in relation to the existing discrete choice and statistics literature. Next, we showed in Section 3.3 how one can leverage the logit-type model formulation to extend existing, asymmetric choice models to the multinomial setting, thereby making such models useful to the transportation community at large. Finally, in Section 3.4, we demonstrated one way to derive entirely new asymmetric choice models based on specific considerations that analysts may have concerning their study. Overall, we presented four new examples of this section's methods by deriving the multinomial clog-log, scobit, uneven logit, and asymmetric logit models. The binary versions of these models are shown in Figure 2 to display the range of shapes that these models embody in comparison to the binary logit model. Note that we display the binary versions of these models instead of their multinomial versions simply for ease of visualization.

In the next Section, we will describe the estimation techniques used in this paper for the logit-type models given by Equation 2. Section 5 will then present the empirical applications of our logit-type models and compare them to the standard multinomial logit model, using the four example models derived in this section. Section 6 will discuss extensions to our work and finally Section 7 will conclude.

## 4. Estimation Techniques

Within transportation, maximum likelihood estimation (MLE) is the most commonly used technique for performing statistical inference on the unknown parameters in one's discrete choice model. For the logit-type models specified in Equation 2, the gradient and hessian of the unknown parameters  $\theta = (\beta, \tau, \gamma)$  can be calculated in closed-form, provided that  $S(\cdot)$  is twice differentiable and provided that the unknown parameters are constrained such that  $S(\cdot)$  exists. The existence of the gradient and hessian permits one to use most numeric optimization methods to try and maximize the likelihood of one's model. Even if  $S(\cdot)$  is not differentiable, one may still be able to make use of sub-gradient methods to perform such numerical maximization.

Despite having closed-form gradients and Hessians, the log-likelihood of one's logit-type model will (in general) not be concave in the unknown parameters  $\theta$ . This lack of concavity can make it difficult to calculate

---

<sup>10</sup>One can show that the binary, asymmetric logit model nests the standard binary logit model when  $\gamma_j = \frac{1}{\|C_i\|}$ .

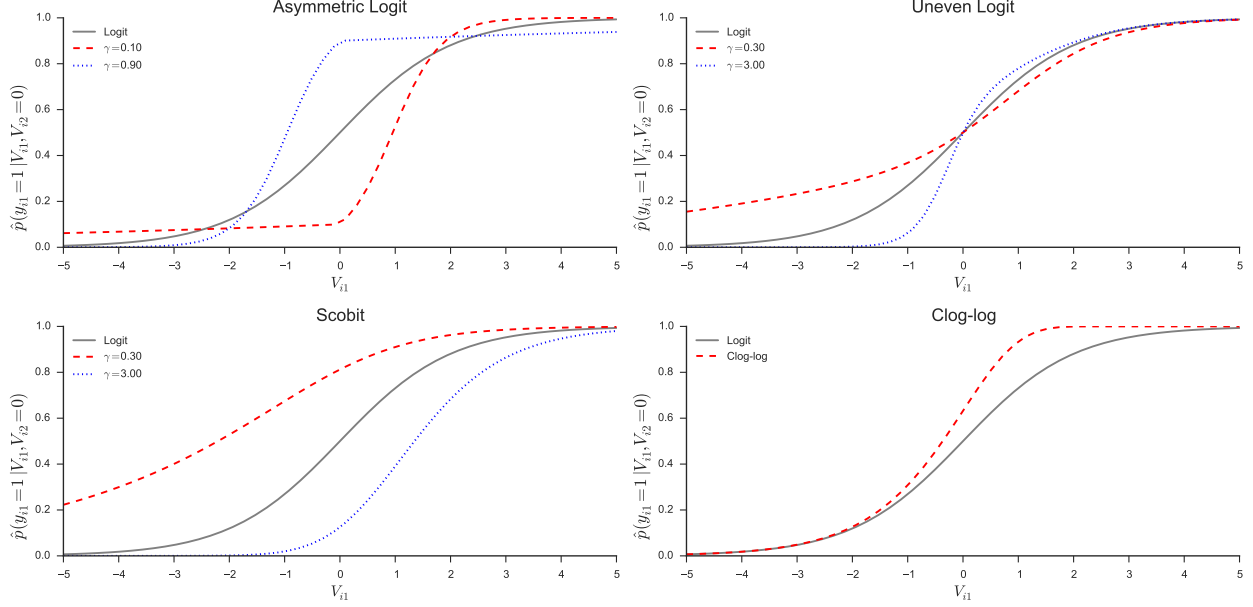


Figure 2: Binary, Asymmetric Choice Models

the MLE estimates for one’s logit-type model. Nevertheless, when possible, we used standard optimization techniques that do not require tuning parameters such as the Newton-Raphson or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms. In cases where standard techniques failed, we resorted to custom-coded gradient descent algorithms. To implement the aforementioned estimation methods, all calculations were carried out using the Python programming language and the NumPy, SciPy, and Pandas packages (McKinney et al., 2010; Van Der Walt et al., 2011; Jones et al., 2014). Moreover, we developed a Python package called PyLogit to perform MLE for the MNL model and the four asymmetric models introduced in Section 3. Our package, PyLogit, is available for public use through the Python Package Index.

## 5. Empirical Applications

This section describes our two policy applications of the asymmetric choice models developed in this paper. These two applications were chosen because they differ in their respective emphases on the aggregate versus disaggregate predictions of the choice models. However, both applications use the same dataset and model specification. In particular, we model the travel mode choice of commuters in the San Francisco Bay Area who are making work or school tours. Our use of a common dataset and model specification allows us to consider the practical differences between the asymmetric and symmetric models based on use case, independent of differences in model inputs.

For our first application, we analyze the impact of a cordon toll in Downtown San Francisco on commute mode shares. As noted in the introduction, commute mode choices are almost always class imbalanced in the US. For instance, as shown in Table 3, approximately 43% of individuals in our sample commute by driving alone while only 5% commute by bicycling. In such class-imbalanced situations, it might be natural to suspect that one’s probability function is asymmetric. We will investigate this hypothesis through statistical tests of our asymmetric choice models versus the MNL model and through each model’s cross-validation performances. To evaluate the possible effects of the asymmetric choice models on policy analyses, in addition to the predictive performance of such models, we investigate the impact of cordon tolls on commute mode choice.

For our second application, we analyze the impact of using our asymmetric choice models in a travel demand management (TDM) setting. In particular, we focus on the use of individualized marketing to increase public transit ridership (Brög, 1998). As a TDM strategy, individualized marketing targets individuals who do not currently use transit but nevertheless could be persuaded to use transit given the information

Table 3: Sample Mode Shares

Travel Mode	Mode Shares (%)
Drive Alone	42.8
Shared Ride-2	15.9
Shared Ride-3+	14.0
Walk-Transit-Walk	10.3
Drive-Transit-Walk	1.5
Walk-Transit-Drive	1.3
Walk	9.4
Bike	4.6

Note: Percentages do not sum to 100 due to rounding error.

and incentives being offered by the marketing campaign (Brög, 1998). An example of one such incentive is the provision of free transit-passes for a limited time. This is the incentive used in our application. By assessing how “switchable” each individual is (Gensch, 1984), choice models such as the MNL model and the asymmetric models developed in this paper are used to select individuals for targeting and transit-pass provision. We then compare the costs and programmatic success of using the MNL model versus our asymmetric logit-type models for target selection in an individualized marketing campaign by treating our sample of individuals as the population of individuals that a transit agency’s pilot marketing program might have access to. Together, the TDM and cordon toll analyses will provide insight into the nature of the practical differences between the asymmetric logit-type models and the traditional MNL model.

In the subsection below, we will describe the dataset used in our two applications. Following this, we will describe the procedures we used for model estimation, model testing, our cordon toll analysis, and our TDM example. Finally, we will report our results, and conclude this section with a discussion.

### 5.1. Data

The data used in this example comes from the 2012 California Household Travel Survey (CHTS). The CHTS was a one day travel diary taken from a stratified sample of households throughout the state of California and portions of Nevada. The complete data collection effort is described in California Department of Transportation (2013). For this study, the overall sample was filtered to include just those individuals commuting to school or work in the San Francisco Bay Area.

Beyond filtering based on geography and trip-purpose, we post-processed the raw CHTS data to construct the final dataset used for model estimation. In particular, we combined the data on individual trips into tours, defined a “chosen travel mode” for each tour, determined the available travel modes for each tour, and assembled the level-of-service variables for each tour. For this study, we used the level-of-service (travel costs, times, and distance) estimates provided by the San Francisco Metropolitan Transportation Commission (MTC). As a result, the set of possible alternatives in our example was defined to be the same as the categories used by MTC. Specifically, eight travel mode alternatives were specified. There were three driving modes, each differentiated by the number of passengers: drive-alone, shared-ride with two passengers, and shared-ride with three or more passengers. There were also three transit modes, each differentiated by their access and egress modes: walk-transit-walk (where walking is used for access and egress), drive-transit-walk, and walk-transit-drive. Finally, there were two non-motorized modes: walking and bicycling. For each tour, the travel mode that was used for the longest distance was used as the “chosen travel mode” for that tour.

After filtering and post-processing, the final dataset consisted of 4,004 home-based work or school tours made by 3,836 individuals (with no individual making more than two tours). The percentage of tours that had their chosen travel mode associated with each of the aforementioned alternatives is shown in Table 3. As mentioned earlier, the proportion of tours associated with each alternative is highly unbalanced, ranging from a low of 1.3% for the share of “walk-transit-drive” tours to a high of 42.8% for drive-alone tours.

### 5.2. Estimation and Testing Procedures

In this subsection, we will describe the procedures we used to perform the estimation, testing, and application of the various logit-type models employed in our example.

### 5.2.1. Estimation

First, to actually perform the numerical optimization necessary for the MLE, the scobit, the uneven logit, and the asymmetric logit models were re-parametrized. In particular, the log-likelihood functions of the scobit and the uneven logit models were expressed in terms of  $\Upsilon_j = \ln(\gamma_j) \forall j$ , and the log-likelihood function of the asymmetric logit model was expressed in terms of  $\Phi_j$  where  $\gamma_j = \frac{\exp(\Phi_j)}{\sum_k \exp(\Phi_k)} \forall j$ . These re-parametrizations allowed for unconstrained optimization of  $\Upsilon_j$  and  $\Phi_j$ , and it led to better estimation results when compared to performing constrained optimizations on the original  $\gamma_j$ 's. Accordingly, our shape parameter estimates for the scobit, uneven logit, and asymmetric logit models are presented in terms of  $\Upsilon_j$  and  $\Phi_j$ , respectively.

### 5.2.2. Testing

In our application, we use two types of model-testing or comparison procedures. First we use “in-sample” testing and comparison where the same sample that is used to estimate our models is then used to compare one model against another. The second type of model comparison and testing procedures that we use is “out-of-sample” where we use one subset of observations to estimate our models and then test our models against a different subset of observations. Because the MNL model is a restricted version of the uneven logit, asymmetric logit, and scobit models, we use log-likelihood ratio tests as our in-sample tests to compare the MNL versus the uneven logit model, the MNL versus the asymmetric logit model, and the MNL versus the scobit model. For our out-of-sample comparisons, we compare all of the models against one-another using ten-fold, stratified cross-validation. For this technique, we separate our data into ten stratified random subsets<sup>11</sup>. Then we iterate through the ten subsets, one at a time, using the selected subset for testing and the other nine subsets for estimation. The models are then compared on the basis of their average log-likelihoods across the ten subsets used for testing.

### 5.2.3. Cordon Toll Analysis

The current congestion pricing proposal for the City of San Francisco is “The Mobility, Access, and Pricing Study” (MAPS) being conducted by the San Francisco County Transportation Authority (2010). The main congestion pricing alternative being studied is a \$3 toll that would be collected from cars passing into or out of the “Northeast Cordon” shown in Figure 3 during the AM peak (6AM - 10AM) or PM peak (3PM - 7PM), with individuals being charged no more than twice per day.

To study the effects of the proposed and similar congestion pricing schemes, we use sample enumeration based on the individual-level sample weights supplied by the CHTS. In particular, we varied the toll amount per crossing, from \$0 to \$5 in \$0.50 increments, calculated the probability of each travel mode for each tour given the current toll amount per crossing, and then used the sample weights to calculate the expected amount of tours using each mode. Care was taken to ensure that we properly calculated if, when, and how many times a tour would result in an individual driving into, out-of, or within the Northeast Cordon so that the toll could be applied as it is has been described in the MAPS study.

Moreover, while it is unlikely that individuals will be using the walk-transit-drive or drive-transit-walk mode to commute into or out-of Downtown San Francisco (due to the lack of public parking lots at subway stations within the Northeast Cordon), we also applied the toll to those modes for people whose destination or origin (respectively) was within the cordon. Our rationale is that the purpose of the toll is to ease congestion within the Northeast Cordon. Using the walk-transit-drive or drive-transit-walk modes to commute into or out-of the Northeast Cordon is not supportive of such a purpose, even if one may not physically drive one's vehicle across the cordon. Our analysis therefore assumes that the agency implementing the congestion charge will devise a way to track and charge individuals using walk-transit-drive or drive-transit-walk to commute into or out-of locations inside the Northeast Cordon.

### 5.2.4. TDM Analysis—Individualized Marketing

To better understand the differences between the standard MNL model and the asymmetric logit-type models developed in this paper, we asked the following two questions. Given a fixed budget to be spent on

---

<sup>11</sup>Stratification is used so that the proportions of tours associated with each travel mode are relatively constant across the subsets.



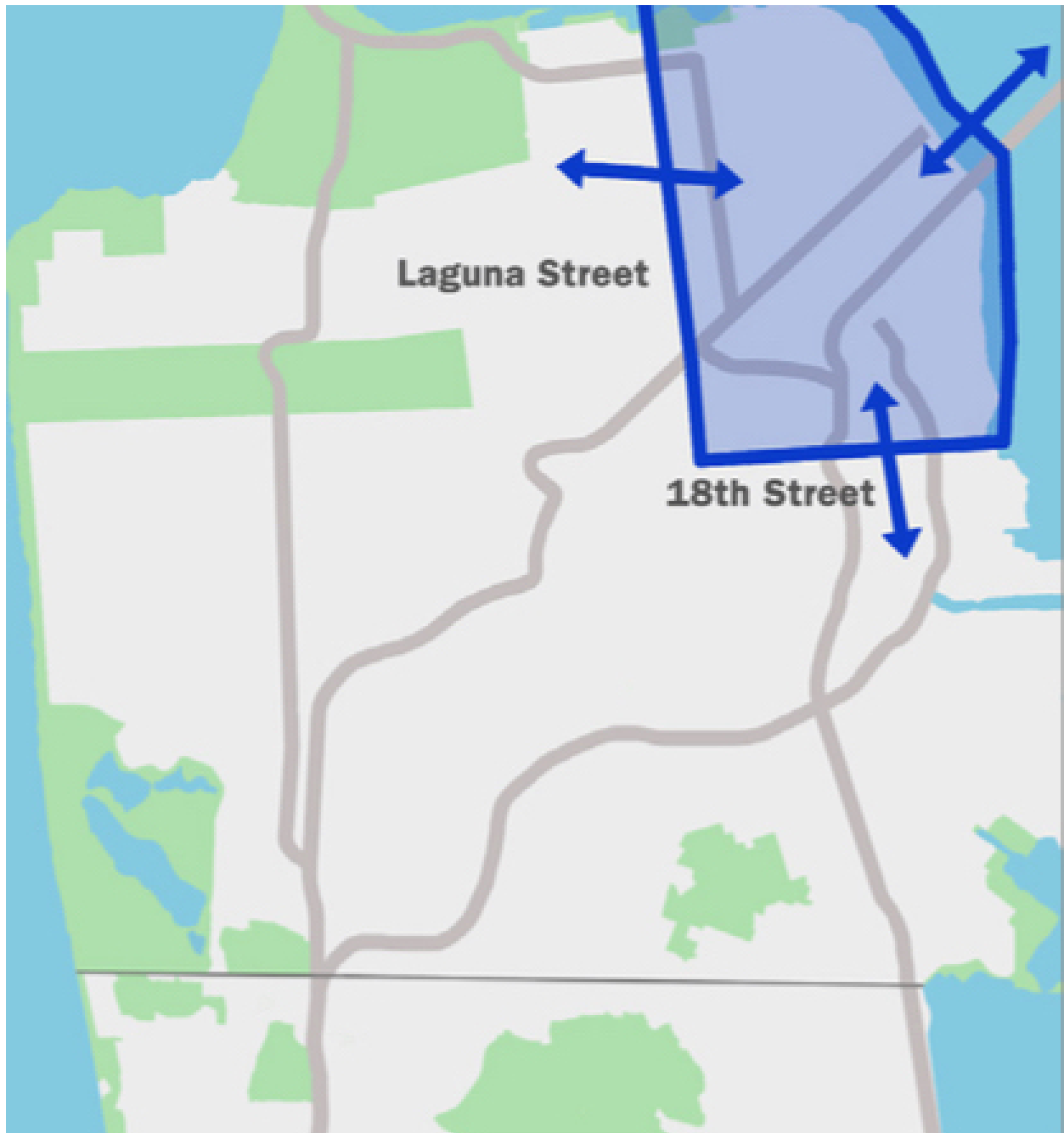


Figure 3: Northeast Cordon for San Francisco Congestion Pricing (San Francisco County Transportation Authority, 2010)

the provision of month-long free transit passes (such as for an individualized marketing pilot program), how would using the various asymmetric choice models for target selection compare to using the MNL model in terms of the dollars spent per expected new transit-rider?

To answer these questions, we needed:

- a way to calculate the costs of transit-pass provision for each targeted individual,
- a way to select individuals for targeting given the choice model being used, and
- a way to assess each targeted individual’s change in the probability of transit usage, given free transit.

First, we calculated the total cost of transit-pass provision for each individual by multiplying each individual’s cost of the “walk-transit-walk” mode by an assumed 22 working days per month. Although one might typically take the cost of a month-long transit pass from relevant transit agencies and use this as the cost of transit-pass provision for each individual, transit agencies in the San Francisco Bay Area such as the Bay Area Rapid Transit (BART) System and Caltrain use distance-based fares. As a result, such agencies do not offer monthly passes, and we based our cost calculations on the individualized transit costs instead of using a single cost for all individuals. The idea is that each individual would be provided with a transit-pass that has been preloaded with the amount of money that is deemed necessary for the individual to complete one walk-transit-walk commute tour per working day for a month.

Secondly, to select individuals for targeting, we assumed the role of an agency that was interested in (1) incentivizing individuals to use the “walk-transit-walk” mode and (2) maximizing the increase in the expected number of walk-transit-walk riders per dollars expended. Based on these goals, our target selection procedure was as follows. We first calculate the probability of using the walk-transit-walk mode with and without a transit pass. Note that the provision of a transit pass would completely eliminate the cost of the walk-transit-walk mode, but it would also reduce the cost of the walk-transit-drive and drive-transit-walk modes by however much the individual would pay in walk-transit-walk costs<sup>12</sup>. Next we divide the change in the walk-transit-walk probability by the total cost of transit provision for each person. Finally, we place the individuals in descending order according to their change in walk-transit-walk probabilities per dollar spent, and we select all individuals from the top of the list such that the total cost of the transit-pass provision for all selected individuals is less than our specified budget. We repeated our analysis for a range of different budgets (\$5,000 - \$60,000) to better understand how the models perform in different scenarios.

Lastly, to assess each targeted individual’s change in the probability of transit usage, we had to choose a model to treat as “truth.” As shown in the upcoming results subsection, the uneven logit model had the best in-sample and out-of-sample log-likelihoods. Given the dominant performance of the uneven logit model, we treated it as the “true” model that would be used to calculate the probability of an individual taking transit with or without a free transit-pass. Each model’s probability predictions were therefore used to select the individuals for targeting as described in the last paragraph, but the uneven logit model was used when assessing the ratio of the total cost of the individualized marketing program to the total increase in the expected number of walk-transit-walk riders.

### 5.3. Results

In this sub-section, we report the results of our model estimation efforts for the standard MNL model and the four asymmetric choice models derived in this paper—the multinomial uneven logit, scobit, asymmetric logit, and clog-log models. The parameter estimates are displayed in Table 4. Note that when interpreting the parameter estimation results displayed in Table 4, one may be inclined to think that the uneven logit model is unidentified or has some sort of error regarding its parameter estimates or standard errors. This is not the case. The t-statistics that are reported are not actually zero, but only appear to be when rounded to three decimal places. Moreover, the log-likelihood function is not completely flat as with an unidentified model. The log-likelihood function of the uneven logit model has a very low (but existent) curvature near the parameter estimates. It is well known in the statistics literature, that if one is estimating both the shape parameters and index coefficients of a parametric link function (i.e. a probability function with parameters that control its shape), then the variance of one’s estimates will be high when the shape parameters and index

---

<sup>12</sup>Of course, we assumed a minimum cost of \$0.

coefficients are highly correlated (Stukel, 1988; Taylor, 1988; Czado and Santner, 1992b). The extremely large standard errors that are present in the uneven logit model are a consequence of the high correlations between its shape parameters ( $\gamma$ ) and index coefficients ( $\beta$ ). As put by Czado, this is the “cost” of estimating the shape of the probability function in addition to the index coefficients (Czado and Santner, 1992b).

In addition to our parameter estimates, we also report the in-sample and out-of-sample predictive performance (i.e. the log-likelihoods) of each of the models in Tables 4, 5, and 6. It can be seen that the asymmetric models generally had better predictive ability than the MNL model, both in-sample and out-of-sample. Finally, beyond the measures of statistical fit, we report the results of our applications on cordon pricing and individualized marketing for a public transportation TDM measure.

Specifically, for the cordon toll analysis, we report the aggregate, automobile-based mode share predictions by each choice model, in relation to the different toll amounts. These aggregate predictions are shown in Figure 4. Moreover, we present a comparison of the disaggregate probability forecasts of the MNL versus the uneven logit model in Figure 5 to highlight the disagreement between the asymmetric models and the MNL model. The maps displayed in Figures 6 and 7 further emphasize the practical significance of the differences between the MNL model and the asymmetric choice models used in this paper.

Finally, for the individualized marketing application, our main results are shown in Figure 8. Defining program efficiency as the total cost of providing the free transit-passes divided by the change in the expected number of walk-transit-walk riders, Figure 8 displays the program efficiencies that are achieved by using each of the choice models for target identification. As mentioned in the following discussion section, it is useful to have Table 5 to help assess the program efficiency results shown in Figure 8. Table 5 decomposes the overall in-sample log-likelihoods achieved by each model into the in-sample log-likelihoods achieved on each travel mode. It allows us to compare the program efficiency results to the predictive ability of each model on specific travel modes instead of just interpreting the program efficiency results based on overall model performance. In general, all of the results mentioned above will be discussed more thoroughly in the discussion section to follow (Section 5.4).

Table 4: MLE Parameter Estimation Results

None	Standard Logit		Uneven Logit		Scobit		Asymmetric Logit		Clog-log	
	Est.	T-stat.	Est.	T-stat.	Est.	T-stat.	Est.	T-stat.	Est.	T-stat.
Alternative Specific Constants										
Shared Ride: 2	-1.010*	-2.079	-0.806**	-3.088	-0.280	-0.735	-1.241**	-3.944	0.969	1.783
Shared Ride: 3+	3.462**	3.254	0.442	1.567	2.596**	3.490	-0.723	-1.725	6.316**	5.477
Walk-Transit-Walk	-0.392	-1.360	0.350	1.535	11.524	0.561	0.489	0.134	-1.741**	-5.711
Drive-Transit-Walk	-2.622**	-8.660	-3.002**	-8.202	4.388	0.303	0.442	0.411	-4.001**	-12.506
Walk-Transit-Drive	-2.977**	-9.725	-3.686**	-9.516	2.566	0.202	0.451	0.362	-4.345**	-13.418
Walk	1.554**	5.101	1.626**	2.756	0.156	0.247	0.850**	11.007	-0.117	-0.346
Bike	-1.106**	-3.628	-0.957**	-2.860	-2.669**	-4.077	0.179	1.371	-2.903**	-8.600
Travel Time, units:min										
All Auto Modes	-0.076**	-13.728	-2.602e-06	0.000	-0.046**	-3.221	-0.042**	-16.376	-0.078**	-13.825
All Transit Modes	-0.027**	-12.768	-375.368	0.000	-0.003	-0.732	-0.016**	-13.642	-0.026**	-12.235
Travel Cost										
Units:\$ All Transit Modes	-0.127**	-3.472	-1,772.138	0.000	-0.015	-0.766	-0.080**	-4.250	-0.210**	-5.436
Units:\$/mi Drive Alone	-5.061**	-3.675	-2.211e-04	0.000	-4.701**	-2.753	-2.464**	-3.890	-10.955**	-7.033
Units:\$/mi SharedRide-2	-20.319**	-4.467	-0.001	0.000	-11.941**	-3.142	-7.863**	-3.987	-47.736**	-9.527
Units:\$/mi SharedRide-3+	-90.922**	-6.165	-0.001	0.000	-32.494**	-3.165	-16.543**	-3.194	-141.947**	-8.877
Travel Distance, units:mi										
Walk	-1.027**	-20.437	-0.852	-0.538	-2.090	-1.585	-0.443**	-13.563	-0.982**	-19.155
Bike	-0.287**	-11.896	-0.211	-0.777	-0.465*	-2.495	-0.165**	-12.729	-0.263**	-11.203
Systematic Heterogeneity										
Autos per licensed drivers (All Auto Modes)	1.213**	9.408	3.689e-05	0.000	0.597**	3.284	0.452**	9.525	0.764**	6.964
Cross-Bay Tour (Shared Ride 2 & 3+)	0.928**	2.839	4.662e-05	0.000	0.906**	3.083	0.549**	3.914	1.707**	5.214
Household Size (Shared Ride 2 & 3+)	0.114*	2.523	5.633e-06	0.000	0.074*	2.396	0.052**	3.018	0.073	1.650
Number of Kids in Household (Shared Ride 2 & 3+)	0.687**	12.820	2.132e-05	0.000	0.327**	3.521	0.248**	11.246	0.682**	12.688
Shape Parameters										
Drive Alone	-	-	10.236	0.001	0.503	1.351	-	-	-	-
Shared Ride: 2	-	-	10.520	0.001	0.804*	2.142	2.007**	3.680	-	-
Shared Ride: 3+	-	-	10.710	0.001	0.987**	2.891	2.805**	5.081	-	-
Walk-Transit-Walk	-	-	-9.408	0.000	2.917	1.856	-1.339	-0.188	-	-
Drive-Transit-Walk	-	-	-9.759	0.000	2.565	1.632	-3.582	-1.686	-	-
Walk-Transit-Drive	-	-	-9.874	0.000	2.434	1.549	-3.953	-1.613	-	-
Walk	-	-	0.146	0.082	-0.811	-1.198	-0.959**	-2.635	-	-
Bicycle	-	-	0.279	0.215	-0.662	-1.300	-1.554**	-3.873	-	-
Log-Likelihood Ratio Stat.										
Log-Likelihood	-5,073.428	-	-4,868.353	-	-4,902.791	-	-4,941.039	-	-5,116.066	-

Note: \* means p-value &lt; 0.05 and \*\* means p-value &lt; 0.01.

Table 5: MLE In-Sample Log-likelihoods by Travel Mode and by Model

	Standard Logit	Uneven Logit	Scobit	Asymmetric Logit	Clog-log
Drive Alone	-1,084.14	-1,045.69	-1,040.07	-1,045.04	-1,092.69
Shared Ride: 2	-1,183.66	-1,137.98	-1,144.92	-1,151.33	-1,196.55
Shared Ride: 3+	-905.80	-826.16	-844.65	-847.51	-926.98
Walk-Transit-Walk	-572.69	-566.88	-569.84	-572.84	-581.04
Drive-Transit-Walk	-184.99	-177.76	-177.20	-182.36	-185.73
Walk-Transit-Drive	-176.93	-167.30	-167.32	-176.53	-175.32
Walk	-520.07	-502.28	-515.38	-519.55	-513.11
Bike	-445.16	-444.31	-443.42	-445.88	-444.65
Total	-5,073.43	-4,868.35	-4,902.79	-4,941.04	-5,116.07

Table 6: MLE Average Out-of-Sample Log-likelihood During 10-fold Cross-Validation

Model	Log-Likelihood
Uneven Logit	-490.12
Scobit	-494.04
Asymmetric Logit	-498.44
Standard Logit	-510.28
Clog-log	-514.63

#### 5.4. Discussion

##### 5.4.1. Model Estimation and Testing

As shown in Tables 4 and 6, the multinomial clog-log model did not perform well relative to the MNL model. However, all of the asymmetric choice models with flexible shapes (i.e. with shape parameters) more accurately predicted the mode choice of individuals in our sample than the MNL model. In particular, there were large differences in in-sample log-likelihoods between the asymmetric choice models with flexible shapes and the MNL model. These differences range from about 132 for the asymmetric logit model to 205 for the uneven logit model. Since all three of the asymmetric choice models with shape parameters nest the MNL model, log-likelihood ratio tests were used to assess whether the differences in model fit were statistically significant. Table 4 shows that all three of the asymmetric, flexible shape models had log-likelihood ratio statistics that were significant at the 0.01 alpha-level. This suggests that the MNL model is inappropriate for this dataset, relative to the flexible, asymmetric choice models used in this paper. Moreover, the greater predictive ability of the uneven logit, the asymmetric logit, and the scobit model was not limited to just the in-sample predictions. The out-of-sample predictions in Table 6 showed exactly the same trends indicated by the in-sample results. Here, the differences in the average out-of-sample log-likelihood during cross-validation ranged from approximately 12 for the asymmetric logit to 20 for the uneven logit. Given that the testing sets in each fold of the cross-validation are about one-tenth the size of the overall dataset, these results are consistent with the in-sample results. This indicates that the greater predictive ability of the flexible, asymmetric choice models as compared to the MNL are real and not due to over-fitting.

##### 5.4.2. Cordon Toll Analysis

In addition to judging whether the improvements offered by one model over another are “statistically significant,” it is important to assess whether such improvements are “practically significant.” One way we assessed the practical impacts of the asymmetric choice models derived in this paper was to conduct an analysis of the effects of a congestion toll in Downtown San Francisco.

At the most basic level, we compare the MNL model and the asymmetric choice models on the basis of their aggregate, predicted mode shares for automobile-based modes (drive alone, shared ride with two passengers, and shared ride with three or more passengers) under various cordon toll charges. Given that the purpose of the congestion toll is to reduce the use of automobile-based modes at peak commute times,

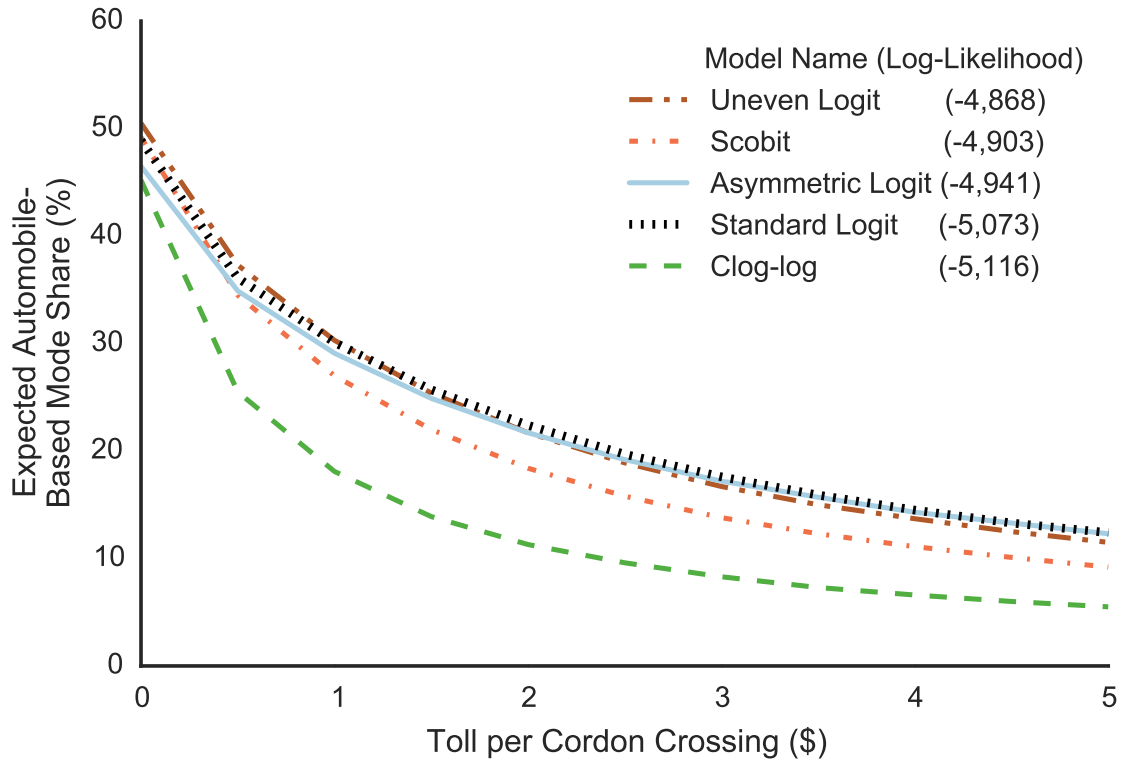


Figure 4: Automobile-Based Mode Shares by Model and by Cordon Toll Amount

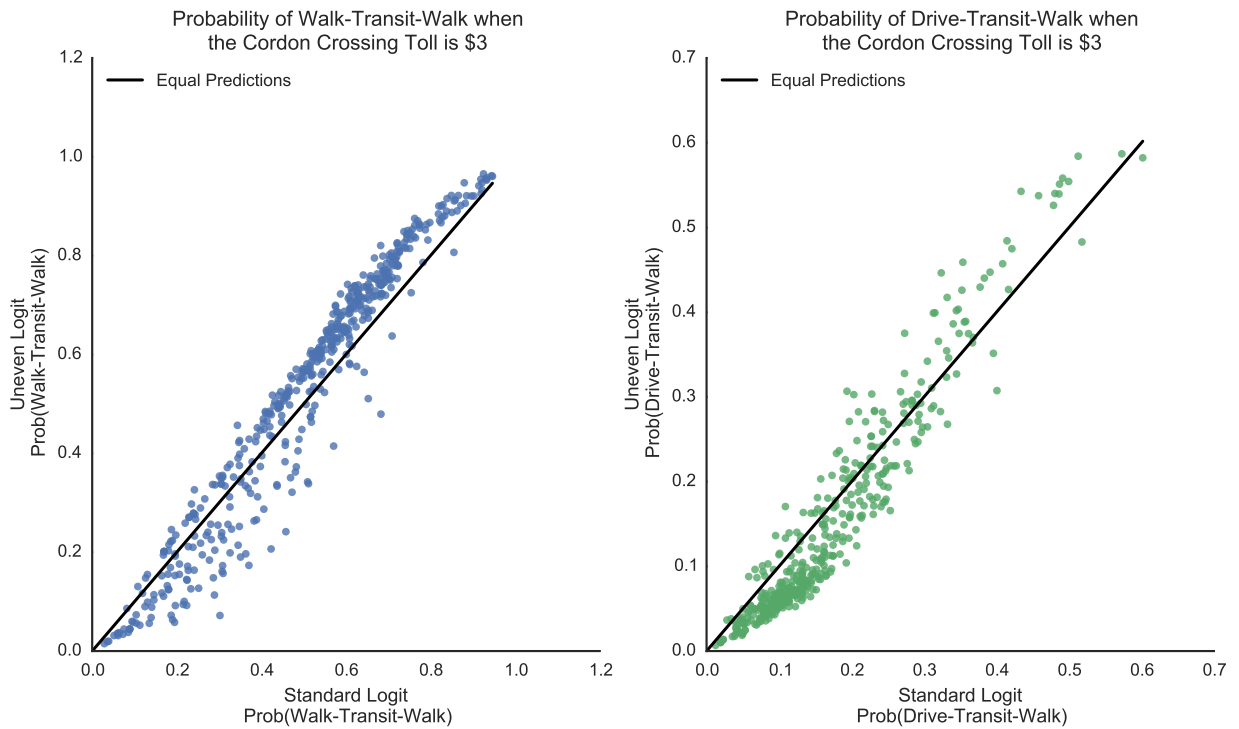


Figure 5: Disaggregate Probability Predictions for Walk-Transit-Walk and Drive-Transit-Walk for the Uneven Logit and the MNL Models

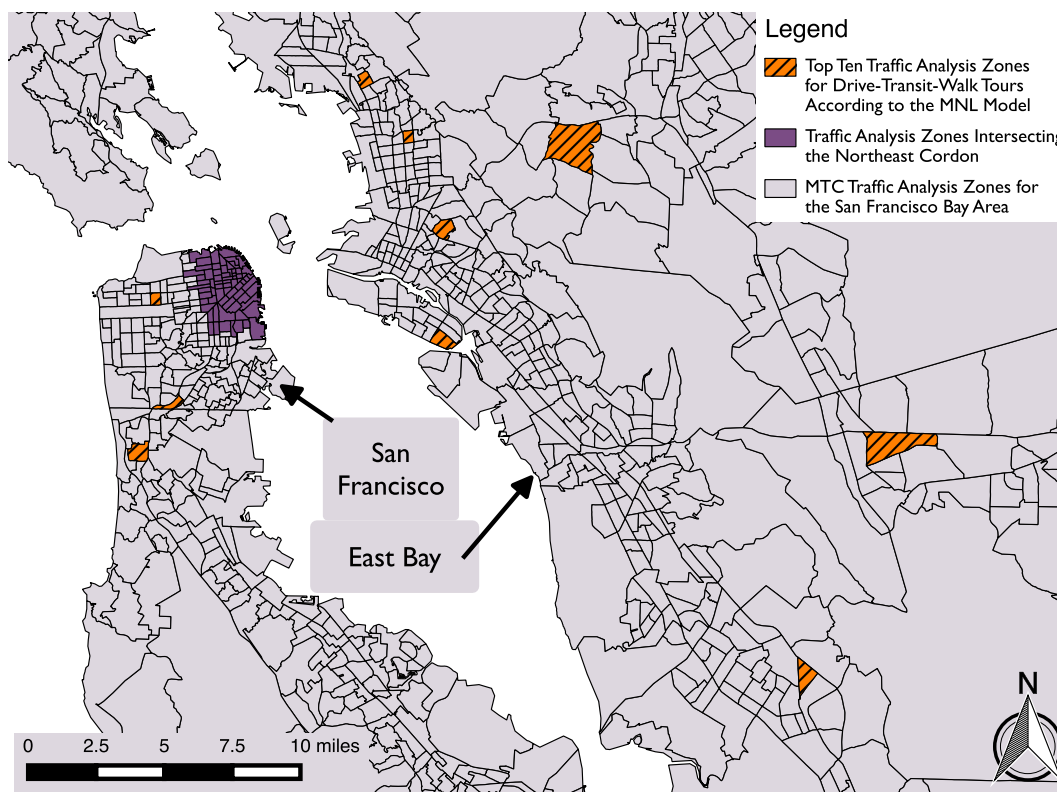


Figure 6: Top Ten Traffic Analysis Zones Producing Drive-Transit-Tours  
According to the MNL Model at \$3 per Cordon Crossing

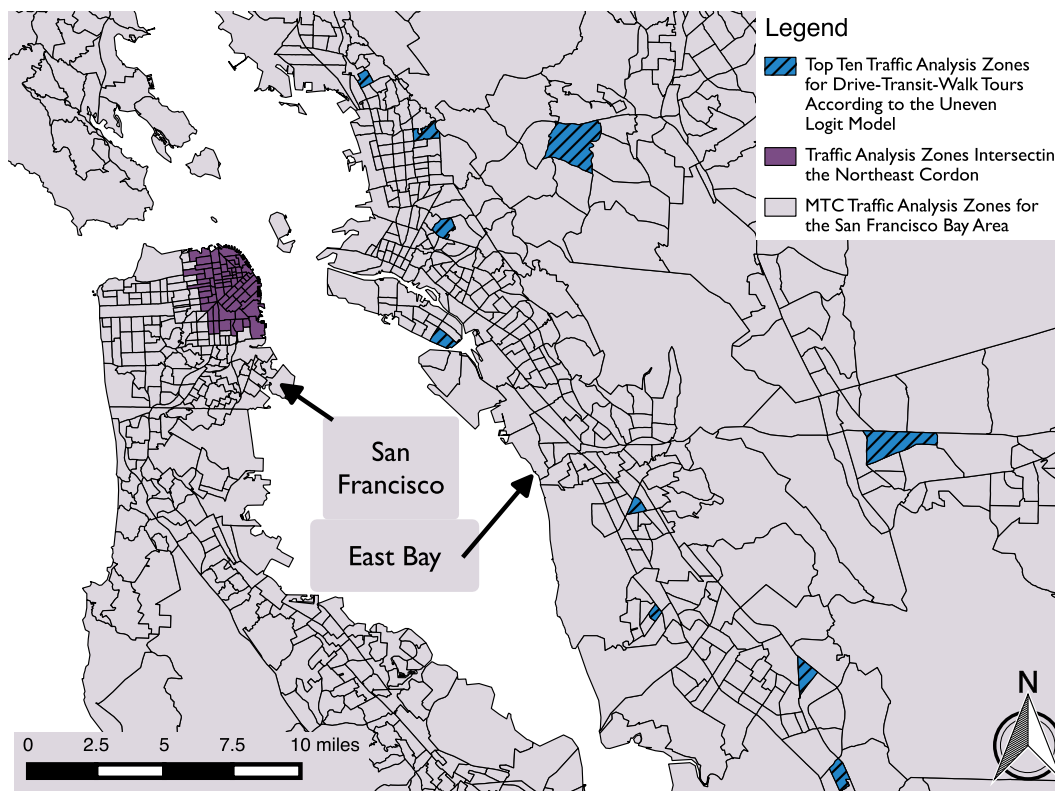


Figure 7: Top Ten Traffic Analysis Zones Producing Drive-Transit-Tours According to the Uneven Logit Model at \$3 per Cordon Crossing

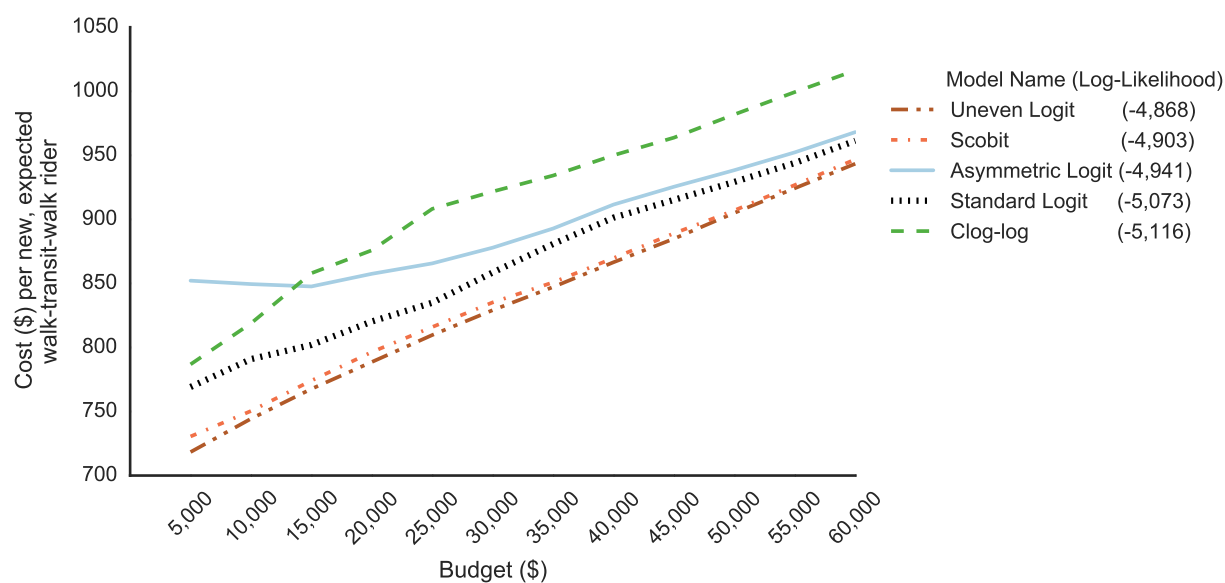


Figure 8: Total Individualized Marketing Costs Per New, Expected Walk-Transit-Walk Rider by Model



large differences in predicted mode share for automobile-based modes would have great ramifications for support and expectations of the congestion tolling scheme. As shown in Figure 4, the aggregate mode share predictions for the automobile based modes, for tours that cross the Northeast Cordon, follows the same general trend for both the MNL and the flexible, asymmetric choice models. Moreover, the differences in the predicted mode shares are minimal. Compared to the flexible, asymmetric models, the MNL model overestimates the mode share of automobile-based modes by 1-3.8% at the SFCTA’s proposed toll of \$3 per cordon crossing, depending on which model is being looked at. However, the MNL and flexible asymmetric models all predicted overall decreases of approximately 31-35% in automobile based mode shares from a \$0 toll to a \$3 toll. In light of the overall predicted mode share changes, the differences between models seems mostly inconsequential from a general planning perspective.

Beyond the basic question of how the aggregate, automobile-based mode shares will change as a result of tolling, a host of disaggregate outputs from mode choice models may be useful for transportation agencies implementing the congestion toll. In particular, to support individuals making their commute trips under the tolling scheme, transportation agencies should make switching to more sustainable modes (such as public transit) as easy and safe as possible. For example, at transit stations where one expects the average number of drive-transit-walk commuters to increase, and where parking capacity is nearly full at peak hours, transit agencies might want to increase parking capacity so that ‘park-and-ride’ trips can be more readily accommodated. However, such actions require knowledge of which transit stations have catchment areas that are going to see large increases in their drive-transit-walk mode shares.

To be accurate, these station-level determinations require accurate predictions of the disaggregate drive-transit-walk probabilities for individuals. In our application, we find substantive disagreements between the MNL model and the flexible, asymmetric choice models. For example, Figure 5 shows the predicted probabilities of walk-transit-walk and drive-transit-walk with a cordon toll of \$3 according to both the MNL model and the uneven logit model for the 4,004 tours in our sample. As can be seen, many of these predicted probabilities disagree. These disagreements are not just an artifact of the \$3 toll, but they exist at every tolling amount we tested, including the base case scenario with no toll. The substantive impact of these individual-level disagreements is that practitioners deciding where to install pedestrian improvements and increase parking capacity based on the MNL model might make misguided decisions: installing infrastructure where it is not needed, or failing to install infrastructure where it is needed. For instance, Figure 6 shows the ten traffic-analysis-zones producing the greatest expected numbers of drive-transit-walk tours into the cordon area for the MNL model at \$3 per cordon crossing, and Figure 7 shows the same for the uneven logit model. As shown by the maps, the MNL model under-predicts the amount of drive-transit-walk trips from the East Bay into the cordon area, relative to the uneven logit model. Practitioners using the MNL model as opposed to the uneven logit model might then incorrectly underestimate the need for increased parking capacity at BART stations in the East Bay, thereby hampering the success of the congestion pricing effort.

#### 5.4.3. *TDM Analysis—Individualized Marketing*

Continuing with the emphasis on disaggregate model differences, this subsection discusses the practical differences for an individualized marketing campaign for TDM. Here, we assume the role of an agency interested in maximizing the increase in the expected number of walk-transit-walk riders per dollars expended. As such, the differences that we are concerned with in this application result from selecting individuals for targeting using each of the choice models being compared in this paper. To the extent that the different models select different individuals, the costs of providing the transit-passes will differ, and the change in the expected number of new walk-transit-walk commuters will differ.

Using the uneven logit model to estimate the “true” change in the expected number of new walk-transit-walk commuters (since the uneven logit model had the highest in-sample and out-of-sample log-likelihoods—see Tables 4 and 6), Figure 8 shows the ratio of the “program efficiencies” achieved by each model, for a range of budgets for purchasing the transit passes. From this Figure, a few insights can be gleaned.

First, when only a small proportion of the sample can be targeted (i.e. when the budget is low), the scobit and uneven logit models make the best uses of money relative to the MNL model. For instance, with a \$5,000 budget, the MNL spends \$770 per new expected walk-transit-walk passenger, while the scobit and uneven logit models spend \$731 and \$719, respectively. If the number of individuals in the marketing program increases while the proportion that is targeted remains the same, such differences in program efficiency will lead to large differences in the number of new walk-transit-walk riders that are attracted using

each model’s targeting list. Second, as the budget increases and the proportion of individuals that can be targeted increases, the differences between the program efficiencies of each model are greatly reduced. This is to be expected. At the limit, there will be a large enough budget to select all individuals for targeting, thus the program costs and the “true” increase in the expected number of walk-transit-walk passengers will be equal across models. Lastly, the ranking of program efficiencies across models depended not on the overall predictive ability of one’s multinomial model but mostly on the predictive ability of one’s model for the travel mode of interest (walk-transit-walk in this case). For instance, since the asymmetric logit model’s in-sample and out-of-sample log-likelihoods are higher than those of the MNL model (see Tables 4 and 6), one would expect the asymmetric logit model to make better targeting selections than the MNL model. However, when one looks at the log-likelihoods of each model for just the walk-transit-walk mode (shown in Table 5), we can see that the asymmetric logit model is actually a worse predictor of the walk-transit-walk mode than the MNL model, even though it has a higher log-likelihood overall. Its program efficiency is therefore worse than that of the MNL model. Another seemingly anomalous fact is that when the budget is low, the clog-log model is able to better target individuals than the asymmetric logit model. This merely reflects the fact that for this sample and relative to the asymmetric logit model, the clog-log model is better able to find the small handful of individuals providing the highest increase in their probability of commuting via walk-transit-walk per dollar spent. However, as the budget increases and the number of individuals that is to be targeted increases, the ranking of program efficiencies return to the predictable state of mimicking the in-sample, log-likelihood rankings for the walk-transit-walk mode.

Overall, for our individualized marketing application, we find that when resources are limited (i.e. when only a small percentage of one’s population can be targeted for marketing), the use of the MNL model can be inefficient as compared to the asymmetric choice models such as the uneven logit and scobit models. In our example, such inefficiencies cost the MNL an additional \$51 per new expected walk-transit-walk rider when compared to the uneven logit model. As the budget for the marketing campaign and the percentage of individuals that could be targeted increased, the disaggregate predictive abilities of each model became less important, and as with the cordon toll application, the practical differences between models became minimal.

### 5.5. Summary

Through our analysis of the commute mode choices of San Francisco Bay Area residents, we found that the three asymmetric choice models with flexible shapes (i.e. those with shape parameters) had much better predictive ability (overall) than the standard MNL model. This result was observed in both in-sample and out-of-sample log-likelihoods. Moreover, these results were corroborated through our log-likelihood ratio test results. All of our flexible asymmetric models had log-likelihoods that were higher than the MNL model, at statistically significant levels.

With regard to practically significant differences, we find that the MNL model and the flexible, asymmetric choice models yield similar *aggregate* inferences in our cordon toll analysis. In particular, results concerning the aggregate mode shares of automobile-based modes at different tolling levels are virtually equal across the different models. The practical differences between the MNL and flexible, asymmetric models comes from the *disaggregate* predictions of the various models for each individual, and the fact that the predictions for some modes may differ greatly. Specifically, the predictions for walk-transit-walk and drive-transit-walk differ greatly between the MNL model and the flexible, asymmetric models in our example.

The practical significance of these differences for our cordon toll application is that discordant inferences are obtained regarding where transit-serving parking supply should be increased. The MNL model suggests that transit-serving parking should be added in San Francisco itself, whereas the asymmetric models imply that the most important places to increase transit-serving parking supply are all in the East Bay. Due to the higher land values in San Francisco and a lower supply of land to devote to parking, the use of the MNL model instead of its better performing, asymmetric counterparts would lead transportation agencies to misguidedly spend much more money providing parking in San Francisco, when the East Bay is likely in greater need for transit-serving parking under a congestion tolling scheme.

For our TDM application, the practical significance of our models is that the asymmetric choice models that predict the walk-transit-walk mode better than the MNL model can better guide investment of the money that is available for individualized marketing of public transit. Specifically, we found that in our

example, when the budget for providing free transit passes was low (\$5,000), the cost of acquiring each new walk-transit-walk rider could be reduced by approximately \$50 and \$40, respectively, by using the uneven logit model and the scobit model for target selection instead of the MNL model. Conversely, as the budget and the percentage of individuals who could be targeted increased, the differences in the disaggregate predictions of the models mattered less and less for target selection and the resulting efficiency of the marketing campaign. This further underscores the fact that the practical usefulness of asymmetric choice models appear to be highest when accurate, disaggregate predictions are needed.

Moving onto the remaining two sections of this paper, we will now transition from discussing our specific applications to looking more broadly at how our work on asymmetric, closed-form, finite-parameter models of multinomial choice can be extended. Then in Section 7, we will conclude by summarizing the theoretical contributions of this paper, highlighting our empirical results, and raising key research questions from this work that should interest academic scholars and professional analysts.

## 6. Extensions

Thus far, all of the individual models and results that have been shown in this paper have been based on the general formulation of logit-type models given by Equation 2. Despite restricting ourselves to that proposed class of models, at least six extensions or future research directions are immediately apparent. In particular, these ideas for future work can be categorized as either (1) direct extensions of the logit-type models developed in this paper, (2) applications of this paper’s ideas to other models, or (3) investigations of the statistical properties of logit-type models. In the following paragraphs we will detail each of the extensions and future research directions that comprise these categories.

Firstly, the logit-type models given in Equation 2 can avoid the symmetry property of standard MNL models, but because they share the same functional form as the MNL model, they retain other undesired properties such as I.I.A. Accordingly, many of the motivations behind existing extensions to the MNL model remain equally applicable to our proposed class of logit-type models. Here we highlight three such extensions. First, models such as the “Heteroskedastic Logit Model” (Steckel and Vanhonacker, 1988; Recker, 1995; Bhat, 1995) allow the scale parameter to vary across observations, and this effectively allows the shape of the resulting probability function to vary across observations. An analogous extension to logit-type models would be to allow  $\gamma_j$  to vary across individuals, such as by parametrizing it as a function of  $x_{ij}$ . Such parametrizations have been successfully used in a transportation context to improve the fit of binary scobit models (Zhang and Timmermans, 2010; Wu et al., 2012), but this type of extension can be more generally applied to any logit-type model that has shape parameters. Second, the wider class of “multivariate extreme value”<sup>13</sup> models (such as the nested and cross-nested logit) generalizes the MNL model, capturing arbitrary correlations between the utilities of an individual’s various alternatives while still maintaining a closed-form expression (Train, 2009). Logit-type models would benefit from similar extensions. As mentioned in Section 3.2, one way to extend logit-type models to account for correlation between the utilities of one’s alternatives is to specify various “aggregation functions” as described by Mattsson et al. (2014) in conjunction with  $w_{ij} = \exp[\tau_j + S(V_{ij}, \gamma_j)]$ . Lastly, MNL models have been extended using various “mixing distributions” to account for taste heterogeneity in their parameters and to provide realistic substitution and correlation patterns between alternatives (Revelt and Train, 1998). These mixed logit approaches use a MNL “kernel” and allow the  $\beta$  coefficients to be randomly distributed throughout the population. Similar mixing strategies could be followed whereby one used a logit-type model as the kernel and a continuous mixing distribution of  $\beta$ s in the model. If using a discrete mixing distribution, i.e. a Latent Class Choice Model (LCCM), an analogous procedure is to use a logit-type model for the class-specific choice model. Such mixing procedures would allow for much greater flexibility and behavioral realism in our proposed logit-type models.

Beyond the direct extensions already mentioned, future research directions include applying the techniques and concerns of this paper to other choice models. Two such research directions seem immediately promising. First, as noted at the end of the last paragraph, one can consider using a logit-type model as the class-specific choice model in a LCCM. However, this still begs the question of what choice model should be

---

<sup>13</sup>This class was originally referred to as “generalized extreme value” (GEV) models (McFadden, 1980). The name multivariate extreme value was adopted to avoid confusion with the pre-existing generalized extreme value distribution (Jenkinson, 1955).

used as the class-membership model. It is not clear that one would necessarily want the class-membership model to have the symmetry property described in the introduction, so it would be interesting to look at the effects of using an asymmetric, logit-type model as the class-membership model in one’s LCCM. There could be large policy impacts from such a change. For instance, imagine one is interested in growing the market share of a desired market segment, such as a latent class of individuals with a predisposition towards using non-motorized modes of transportation). If that market segment is forecast to grow much more slowly when using an asymmetric model for the class membership probabilities as opposed to a MNL, and the asymmetric model fits one’s data better, then policy-makers may need to take more aggressive measures to increase the market shares of the desired class. Secondly, the logit-type models developed in this paper were based on the desire to make the MNL model asymmetric. However, as stated above, this logit-based lineage leads to the inheritance of the other undesired properties of the logit model such as I.I.A. It would be interesting to instead try and make other, non-logit-based, choice models asymmetric. For instance, the Exponential Choice model is not based on the logit model, yet it shares some of the attractive properties of the logit model. In particular, it has a closed-form probability equation, it has a log-likelihood that is concave with respect to the  $\beta$ s to be estimated, and it does not have the I.I.A. property (Alptekinoglu and Semple, 2016). However, it is a symmetric probability function<sup>14</sup>. It would be quite interesting to develop an asymmetric analogue to the Exponential Choice model, as such a model would avoid both the I.I.A. property and the symmetry property.

Finally, there are a number of statistical questions regarding logit-type models that remain to be investigated. One of these questions is what is the best way to estimate one’s logit-type model? As was noted in Section 4, MLE was sometimes difficult for the four logit-type models derived in this paper. One response is to use Bayesian techniques to estimate the logit-type models since these techniques do not require maximization of an objective function. However, Bayesian estimation techniques can potentially lead to long estimation times, depending on one’s model, dataset, and specific estimation method. It would be useful to investigate the properties of Bayesian and other estimation techniques on logit-type models. For instance, it has already been shown that maximum entropy estimation (Donoso et al., 2011) may be a better estimation technique than MLE for nested logit models. Further research should be done with logit-type models to investigate the implications, the possible equivalences, and the relative merits and drawbacks of various estimation techniques such as bayesian inference, maximum entropy, method of moments, minimum chi-square estimation (Berkson, 1980), etc.

## 7. Conclusion

In this paper’s introduction, we called attention to a symmetry property of common discrete choice models such as the MNL model and the simple probit model. Arguing that it is often undesirable for one’s discrete choice model to a-priori be symmetric, we introduced a class of “logit-type” models that allow one to specify choice models of varying shapes and asymmetries, without entailing restrictions on the sign or magnitude of the index. Essentially, logit-type models replace the index,  $V_{ij} = x_{ij}\beta$  in the MNL model with functions,  $S(\cdot)$ , that depend on the index and a finite number of shape parameters that control the shape of the probability function. By ensuring that this new function is asymmetric with respect to the index, we avoid symmetry in our logit-type models.

Next, we showed that our proposed class of models is both a parametrization of the class of models introduced by Mattsson et al. (2014) as well as a generalization of numerous existing, asymmetric choice models from both the transportation discipline as well as statistics. This nesting of existing models was used to devise a methodology for extending numerous pre-existing models to the “conditional” and multinomial settings. Such extensions greatly increase the number of situations that can be modeled by existing asymmetric choice models and increase the relevance of such models to transportation researchers whom often study inherently multinomial choice contexts. As examples of the proposed method, we extended two existing models—the clog-log model and the scobit model—to the multinomial setting for the first time.

Recognizing that the existing asymmetric choice models may not always suit a researcher’s needs, we proposed a method for creating new, asymmetric choice models. We break from recent trends in transportation

---

<sup>14</sup>This assertion is made based on plotting the choice probabilities for the binary exponential choice model.

whereby one first specifies the distribution of each alternative’s utility to each individual and then derives the choice probability functions as a result. Our paper takes the opposite approach of directly specifying the form of the choice probability functions, knowing that our logit-type models can be derived from innumerable distributions of the utilities. Doing so frees us to specify the choice probability functions according to the properties that we find desirable for our study. To demonstrate our proposed procedure, we derived two new choice models that generalize the MNL model: the asymmetric logit model and the uneven logit model.

To test the four new models derived in this paper against the standard MNL model, we applied all of these models to an analysis of travel mode choice in the San Francisco Bay Area. We find that all of the asymmetric choice models with flexible shapes (i.e. those with shape parameters to be estimated from the data) were able to fit the data better according to both in-sample and out-of-sample log-likelihoods. The difference in fit, for our example, was not just statistically significant but quite dramatic—on the order of more than 200 log-likelihood points for a dataset of only 4,004 individuals with 8 alternatives. Moreover, beyond the statistical fit and predictive ability of the various models, we showed that switching to asymmetric choice models can also entail serious policy implications. When looking at the effects of a cordon toll in Downtown San Francisco, we found that relative to the flexible asymmetric choice models (which had greater predictive power), the MNL model over-predicted the number of drive-transit-walk tours coming from San Francisco. Such over-predictions would encourage transportation agencies to erroneously invest more in increasing transit-serving parking supply in San Francisco as compared to the East Bay, where all of the other asymmetric models predict high expected numbers of drive-transit-walk tours. Moreover, in our TDM application, we find that the uneven logit model and the scobit model are able to better target individuals for marketing when funding for such a campaign is limited. In particular, the uneven logit and scobit models are able to reduce the cost of acquiring each new walk-transit-walk customer by approximately \$50 to \$40 relative to the MNL model when the marketing budget is only \$5,000. These results suggest that while asymmetric models may not always outperform symmetric ones, asymmetric choice models are at least worth testing in one’s analysis as they might have better statistical performance and entail substantive policy and financial implications.

Lastly, while this paper presents a new class of models as well as four particular instances of this new class, many extensions to this work and future research directions remain. By direct analogy with MNL models, it will be of interest to extend logit-type models to account for arbitrary correlation structures between the various utilities of each alternative. Moreover, it will be interesting to make use of mixture formulations to incorporate taste heterogeneity and flexible patterns of substitution between alternatives. Regarding applications, further investigation remains to be done on the effect of incorporating logit-type models into other contexts (such as modeling market segmentation in LCCMs) and on the effect of incorporating asymmetry into choice models with different functional forms from the logit model (such as the Exponential Choice model). Alongside all of the research directions mentioned above, there will of course be a need to answer statistical questions related to the proposed model-class, including questions of how best to estimate logit-type models and how one can check the appropriateness of a given function,  $S(\cdot)$ , for one’s data.

## Acknowledgements

We wish to thank James A. Goulet for many stimulating conversations in the beginning stages of this research. Additionally, we would like to thank Michael Fratoni for computational assistance in the beginning of this project and Madeleine Sheehan for her constructive criticism of this manuscript. Any errors or omissions are, of course, our own. Lastly, we thank UCCONNECT and the California State Department of Transportation for funding this research effort.

## Appendix

Here we provide the derivation of Equation 11. It is based on Equation 16 and Equation 45 of Buja et al. (2005). Note that in all equations below, we use the notation introduced in Section 3.3.

First, Equation 16 of Buja et al. states that:

$$L_2(\hat{p}(V_{i1})) = \int_0^{\hat{p}(V_{i1})} tw(t) dt \quad (\text{A1})$$

Applying the Fundamental Theorem of Calculus to Equation A1, we can write:

$$\frac{d[L_2(\hat{p}(V_{i1}))]}{d\hat{p}} = \hat{p}(V_{i1}) w(\hat{p}(V_{i1})) \quad (\text{A2})$$

At the same time, Equation 45 of Buja et al. states:

$$1 = w(\hat{p}(V_{i1})) \hat{p}'(V_{i1}) \quad (\text{A3})$$

Assuming that  $\hat{p}'(V_{i1}) \neq 0$ , we can rearrange Equation A3 as follows:

$$\frac{1}{\hat{p}'(V_{i1})} = w(\hat{p}(V_{i1})) \quad (\text{A4})$$

Finally, substituting Equation A4 into Equation A2 yields Equation 11.

## References

## References

- Alptekinoglu, A., Semple, J.H., 2016. The Exponential Choice Model: A New Alternative for Assortment and Price Optimization. *Operations Research* 64, 79–93. URL: <http://pubsonline.informs.org/doi/abs/10.1287/opre.2015.1459>, doi:10.1287/opre.2015.1459.
- Aranda-Ordaz, F.J., 1981. On two families of transformations to additivity for binary response data. *Biometrika* 68, 357–363. URL: <http://biomet.oxfordjournals.org/content/68/2/357>, doi:10.1093/biomet/68.2.357.
- Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Optimization with Sparsity-Inducing Penalties. *Found. Trends Mach. Learn.* 4, 1–106. URL: <http://dx.doi.org/10.1561/22000000015>, doi:10.1561/22000000015.
- Bazán, J.L., Bolfarine, H., Branco, M.D., 2010. A Framework for Skew-Probit Links in Binary Regression. *Communications in Statistics - Theory and Methods* 39, 678–697. URL: <http://dx.doi.org/10.1080/03610920902783849>, doi:10.1080/03610920902783849.
- Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press.
- Berkson, J., 1980. Minimum Chi-Square, not Maximum Likelihood! *The Annals of Statistics* 8, 457–487. URL: <http://www.jstor.org/stable/2240587>.
- Bhat, C.R., 1995. A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological* 29, 471–483. URL: <http://www.sciencedirect.com/science/article/pii/0191261595000156>, doi:10.1016/0191-2615(95)00015-6.
- Bianco, A.M., Yohai, V.J., 1996. Robust Estimation in the Logistic Regression Model, in: Rieder, H. (Ed.), *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Springer New York. number 109 in *Lecture Notes in Statistics*, pp. 17–34. URL: [http://link.springer.com/chapter/10.1007/978-1-4612-2380-1\\_2](http://link.springer.com/chapter/10.1007/978-1-4612-2380-1_2). doi: 10.1007/978-1-4612-2380-1\_2.
- Brög, W., 1998. Individualized marketing: implications for transportation demand management. *Transportation Research Record: Journal of the Transportation Research Board*, 116–121.
- Buja, A., Stuetzle, W., Shen, Y., 2005. Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications,” manuscript, available at [www-stat.wharton.upenn.edu/~buja](http://www-stat.wharton.upenn.edu/~buja).
- Calabrese, R., Osmetti, S.A., 2013. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics* 40, 1172–1188. URL: <http://dx.doi.org/10.1080/02664763.2013.784894>, doi:10.1080/02664763.2013.784894.
- California Department of Transportation, 2013. 2010-2012 California Household Travel Survey Final Report. Technical Report. URL: <http://www.dot.ca.gov/hq/tsip/FinalReport.pdf>.
- Carroll, R.J., Pederson, S., 1993. On Robustness in the Logistic Regression Model. *Journal of the Royal Statistical Society. Series B (Methodological)* 55, 693–706. URL: <http://www.jstor.org/stable/2345881>.
- Castillo, E., Menéndez, J.M., Jiménez, P., Rivas, A., 2008. Closed form expressions for choice probabilities in the Weibull case. *Transportation Research Part B: Methodological* 42, 373–380. URL: <http://www.sciencedirect.com/science/article/pii/S019126150700077X>, doi:10.1016/j.trb.2007.08.002.
- Chen, M.H., Dey, D.K., Shao, Q.M., 1999. A New Skewed Link Model for Dichotomous Quantal Response Data. *Journal of the American Statistical Association* 94, 1172–1186. URL: <http://www.jstor.org/stable/2669933>, doi:10.2307/2669933.

- Chorus, C., van Cranenburgh, S., Dekker, T., 2014. Random regret minimization for consumer choice modeling: Assessment of empirical evidence. *Journal of Business Research* 67, 2428–2436.
- Czado, C., 1992. On Link Selection in Generalized Linear Models, in: Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G. (Eds.), *Advances in GLIM and Statistical Modelling*. Springer New York. number 78 in *Lecture Notes in Statistics*, pp. 60–65. URL: [http://link.springer.com/chapter/10.1007/978-1-4612-2952-0\\_10](http://link.springer.com/chapter/10.1007/978-1-4612-2952-0_10).
- Czado, C., 1994. Parametric link modification of both tails in binary regression. *Statistical Papers* 35, 189–201. URL: <http://link.springer.com/article/10.1007/BF02926413>, doi:10.1007/BF02926413.
- Czado, C., Santner, T.J., 1992a. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* 33, 213–231. URL: <http://www.sciencedirect.com/science/article/pii/0378375892900695>, doi:10.1016/0378-3758(92)90069-5.
- Czado, C., Santner, T.J., 1992b. Orthogonalizing parametric link transformation families in binary regression analysis. *Canadian Journal of Statistics* 20, 51–61.
- Daganzo, C., 1979. *Multinomial Probit: The Theory and Its Application to Demand Forecasting*. Academic Press.
- Das, I., Mukhopadhyay, S., 2014. On generalized multinomial models and joint percentile estimation. *Journal of Statistical Planning and Inference* 145, 190–203. URL: <http://www.sciencedirect.com/science/article/pii/S0378375813002103>, doi:10.1016/j.jspi.2013.08.015.
- Dawid, A.P., 2006. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics* 59, 77–93. URL: <http://link.springer.com/article/10.1007/s10463-006-0099-8>, doi:10.1007/s10463-006-0099-8.
- Donoso, P., De Grange, L., González, F., 2011. A Maximum Entropy Estimator for the Aggregate Hierarchical Logit Model. *Entropy* 13, 1425–1445. URL: <http://www.mdpi.com/1099-4300/13/8/1425>, doi:10.3390/e13081425.
- Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222, 309–368.
- Fosgerau, M., Bierlaire, M., 2009. Discrete choice models with multiplicative error terms. *Transportation Research Part B: Methodological* 43, 494–505. URL: <http://www.sciencedirect.com/science/article/pii/S0191261508001215>, doi:10.1016/j.trb.2008.10.004.
- Gensch, D.H., 1984. Targeting the switchable industrial customer. *Marketing Science* 3, 41–54.
- Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102, 359–378. URL: <http://dx.doi.org/10.1198/016214506000001437>, doi:10.1198/016214506000001437.
- Goleţ, I., 2014. Symmetric and Asymmetric Binary Choice Models for Corporate Bankruptcy. *Procedia - Social and Behavioral Sciences* 124, 282–291. URL: <http://www.sciencedirect.com/science/article/pii/S1877042814020357>, doi:10.1016/j.sbspro.2014.02.487.
- Guerrero, V.M., Johnson, R.A., 1982. Use of the Box-Cox transformation with binary response models. *Biometrika* 69, 309–314. URL: <http://biomet.oxfordjournals.org/content/69/2/309>, doi:10.1093/biomet/69.2.309.
- Härdle, W., Spokoiny, V., Sperlich, S., others, 1997. Semiparametric single index versus fixed link function modelling. *The Annals of Statistics* 25, 212–243. URL: <http://projecteuclid.org/euclid.aos/1034276627>.



- Hennig, C., Kutlukaya, M., 2007. Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal* 5, 19–39. URL: <https://ine.pt/revstat/pdf/rs070102.pdf>.
- Horowitz, J.L., 1993. Semiparametric and nonparametric estimation of quantal response models, Elsevier. volume 11 of *Handbook of Statistics*, pp. 45–72. URL: <http://www.sciencedirect.com/science/article/pii/S0169716105800379>.
- Horowitz, J.L., 2010. *Semiparametric and Nonparametric Methods in Econometrics*. Springer Science & Business Media.
- Jenkinson, A.F., 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society* 81, 158–171. URL: <http://onlinelibrary.wiley.com/doi/10.1002/qj.49708134804/abstract>, doi:10.1002/qj.49708134804.
- Jiang, X., Dey, D.K., Prunier, R., Wilson, A.M., Holsinger, K.E., 2013. A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics* 7, 2180–2204. URL: <http://arxiv.org/abs/1401.1915>, doi:10.1214/13-A0AS663. arXiv: 1401.1915.
- Jones, E., Oliphant, T., Peterson, P., 2014. {SciPy}: open source scientific tools for {Python} .
- Keener, R.W., 2006. *Statistical theory: notes for a course in theoretical statistics*.
- Kim, H.J., 2002. Binary Regression with a Class of Skewed t-Link Models. *Communications in Statistics - Theory and Methods* 31, 1863–1886. URL: <http://dx.doi.org/10.1081/STA-120014917>, doi:10.1081/STA-120014917.
- Kim, S., Chen, M.H., Dey, D.K., 2008. Flexible generalized t-link models for binary response data. *Biometrika* 95, 93–106. URL: <http://biomet.oxfordjournals.org/content/95/1/93>, doi:10.1093/biomet/asm079.
- Koenker, R., Yoon, J., 2009. Parametric links for binary choice models: A Fisherian–Bayesian colloquy. *Journal of Econometrics* 152, 120–130. URL: <http://www.sciencedirect.com/science/article/pii/S0304407609000207>, doi:10.1016/j.jeconom.2009.01.009.
- Komori, O., Eguchi, S., Ikeda, S., Okamura, H., Ichinokawa, M., Nakayama, S., 2015. An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution* , n/a–n/aURL: <http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12473/abstract>, doi:10.1111/2041-210X.12473.
- Kyung, M., Gill, J., Ghosh, M., Casella, G., 2010. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 5, 369–411. URL: <http://projecteuclid.org/euclid.ba/1340218343>.
- Leong, W., Hensher, D.A., 2015. Contrasts of relative advantage maximisation with random utility maximisation and regret minimisation. *Journal of Transport Economics and Policy (JTEP)* 49, 167–186.
- Li, B., 2011. The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis. *Transportation Research Part B: Methodological* 45, 461–473. URL: <http://www.sciencedirect.com/science/article/pii/S0191261510001190>, doi:10.1016/j.trb.2010.09.007.
- Masnadi-shirazi, H., Vasconcelos, N., 2010. Variable margin losses for classifier design, in: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (Eds.), *Advances in Neural Information Processing Systems* 23. Curran Associates, Inc., pp. 1576–1584. URL: <http://papers.nips.cc/paper/4024-variable-margin-losses-for-classifier-design.pdf>.
- Mattsson, L.G., Weibull, J.W., Lindberg, P.O., 2014. Extreme values, invariance and choice probabilities. *Transportation Research Part B: Methodological* 59, 81–95. URL: <http://www.sciencedirect.com/science/article/pii/S0191261513001987>, doi:10.1016/j.trb.2013.10.014.
- McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*, Second Edition. 2 edition ed., Chapman and Hall/CRC, Boca Raton.

- McFadden, D., 1972. Conditional Logit Analysis of Qualitative Choice Behavior. Working Paper Institute of Urban and Regional Development URL: <http://trid.trb.org/view.aspx?id=235187>.
- McFadden, D., 1980. Econometric Models for Probabilistic Choice Among Products. *The Journal of Business* 53, S13–S29. URL: <http://www.jstor.org/stable/2352205>.
- McKinney, W., et al., 2010. Data structures for statistical computing in python, in: *Proceedings of the 9th Python in Science Conference*, pp. 51–56.
- Merkle, E.C., Steyvers, M., 2013. Choosing a Strictly Proper Scoring Rule. *Decision Analysis* 10, 292–304. URL: <http://pubsonline.informs.org/doi/abs/10.1287/deca.2013.0280>, doi:10.1287/deca.2013.0280.
- Nagler, J., 1994. Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science* 38, 230–255. URL: <http://www.jstor.org/stable/2111343>, doi:10.2307/2111343.
- Nakayama, S., Chikaraishi, M., 2015. Unified closed-form expression of logit and weibit and its extension to a transportation network equilibrium assignment. *Transportation Research Part B: Methodological* 81, Part 3, 672–685. URL: <http://www.sciencedirect.com/science/article/pii/S0191261515001666>, doi:10.1016/j.trb.2015.07.019.
- Pregibon, D., 1980. Goodness of Link Tests for Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29, 15–14. URL: <http://www.jstor.org/stable/2346405>, doi:10.2307/2346405.
- Pregibon, D., 1982. Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics* 38, 485–498. URL: <http://www.jstor.org/stable/2530463>, doi:10.2307/2530463.
- Prentice, R.L., 1976. A Generalization of the Probit and Logit Methods for Dose Response Curves. *Biometrics* 32, 761–768. URL: <http://www.jstor.org/stable/2529262>, doi:10.2307/2529262.
- Recker, W.W., 1995. Discrete choice with an oddball alternative. *Transportation Research Part B: Methodological* 29, 201–211. URL: <http://www.sciencedirect.com/science/article/pii/S019126159500002U>, doi:10.1016/0191-2615(95)00002-U.
- Reid, M.D., Williamson, R.C., 2010. Composite binary losses. *The Journal of Machine Learning Research* 11, 2387–2422. URL: <http://dl.acm.org/citation.cfm?id=1953012>.
- Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households’ choices of appliance efficiency level. *Review of economics and statistics* 80, 647–657. URL: <http://www.mitpressjournals.org/doi/abs/10.1162/003465398557735>.
- San Francisco County Transportation Authority, 2010. San Francisco Mobility, Access, and Pricing Study. Technical Report. San Francisco County Transportation Authority. URL: [http://www.sfcta.org/sites/default/files/content/Planning/CongestionPricingFeasibilityStudy/PDFs/MAPS\\_study\\_final\\_lo\\_res.pdf](http://www.sfcta.org/sites/default/files/content/Planning/CongestionPricingFeasibilityStudy/PDFs/MAPS_study_final_lo_res.pdf).
- Scott, C., 2012. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics* 6, 958–992. URL: <http://projecteuclid.org/euclid.ejs/1337951630>, doi:10.1214/12-EJS699.
- Steckel, J.H., Vanhonacker, W.R., 1988. A Heterogeneous Conditional Logit Model of Choice. *Journal of Business & Economic Statistics* 6, 391–398. URL: <http://www.jstor.org/stable/1391892>, doi:10.2307/1391892.
- Stukel, T.A., 1988. Generalized Logistic Models. *Journal of the American Statistical Association* 83, 426–431. URL: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478613>, doi:10.1080/01621459.1988.10478613.
- Taylor, J.M., 1988. The cost of generalizing logistic regression. *Journal of the American Statistical Association* 83, 1078–1083.

- Train, K., 2009. Discrete Choice Methods With Simulation. 2 ed., Cambridge University Press, New York, NY, USA.
- Van Der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering* 13, 22–30.
- Vijverberg, C.P.C., Vijverberg, W.P.M., 2012. Pregibit: A Family of Discrete Choice Models. SSRN Scholarly Paper ID 2010974. Social Science Research Network. Rochester, NY. URL: <http://papers.ssrn.com/abstract=2010974>.
- Vijverberg, W.P.M., 2000. Betit: A Family That Nests Probit and Logit. SSRN Scholarly Paper ID 264789. Social Science Research Network. Rochester, NY. URL: <http://papers.ssrn.com/abstract=264789>.
- Wang, X., Dey, D.K., 2010. Generalized Extreme Value Regression for Binary Response Data: An Application to B@B Electronic Payments System Adoption. *The Annals of Applied Statistics* 4, 2000–2023. URL: <http://www.jstor.org/stable/23362457>.
- Winkler, R.L., 1994. Evaluating Probabilities: Asymmetric Scoring Rules. *Management Science* 40, 1395–1405. URL: <http://www.jstor.org/stable/2632926>.
- Wu, L., Zhang, J., Fujiwara, A., Chikaraishi, M., 2012. Analysis of Tourism Generation Incorporating the Influence of Constraints Based on a Scobit Model. *Asian Transport Studies* 2, 19–33. doi:10.11175/eastsats.2.19.
- Xu, H., Caramanis, C., Mannor, S., 2012. Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 187–193. doi:10.1109/TPAMI.2011.177.
- Yates, F., 1955. The Use of Transformations and Maximum Likelihood in the Analysis of Quantal Experiments Involving Two Treatments. *Biometrika* 42, 382–403. URL: <http://www.jstor.org/stable/2333385>, doi:10.2307/2333385.
- Zhang, J., Timmermans, H., 2010. Scobit-Based Panel Analysis of Multitasking Behavior of Public Transport Users. *Transportation Research Record: Journal of the Transportation Research Board* 2157, 46–53. URL: <http://trrjournalonline.trb.org/doi/abs/10.3141/2157-06>, doi:10.3141/2157-06.
- Zhang, J., Xu, L., Fujiwara, A., 2011. Developing an integrated scobit-based activity participation and time allocation model to explore influence of childcare on women’s time use behaviour. *Transportation* 39, 125–149. URL: <http://link.springer.com/article/10.1007/s11116-011-9321-5>, doi:10.1007/s11116-011-9321-5.